ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746162

PARTIALLY FAKE AUDIO DETECTION BY SELF-ATTENTION-BASED FAKE SPAN DISCOVERY

Haibin Wu[14]*, Heng-Cheng Kuo*[2]*, Naijun Zheng*[5]*, Kuo-Hsuan Hung*[2]
Hung-yi Lee[1]*, Yu Tsao*[2]*, Hsin-Min Wang*[3]*, Helen Meng*[45]

[3] Institute of Information Science, Academia Sinica, Taiwan
[1] Graduate Institute of Communication Engineering, National Taiwan University
[2] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[5] Human-Computer Communications Laboratory, The Chinese University of Hong Kong
[4]Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong

ABSTRACT

The past few years have witnessed the significant advances of speech synthesis and voice conversion technologies. However, such technologies can undermine the robustness of broadly implemented biometric identification models and can be harnessed by in-the-wild attackers for illegal uses. The ASVspoof challenge mainly focuses on synthesized audios by advanced speech synthesis and voice conversion models, and replay attacks. Recently, the first Audio Deep Synthesis Detection challenge (ADD 2022) extends the attack scenarios into more aspects. Also, ADD 2022 is the first challenge to propose the partially fake audio detection task. Such brand new attacks are dangerous and how to tackle such attacks remains an open question. Thus, we propose a novel framework by introducing the question-answering (fake span discovery) strategy with the self-attention mechanism to detect partially fake audios. The proposed fake span detection module tasks the anti-spoofing model to predict the start and end positions of the fake clip within the partially fake audio, address the model's attention into discovering the fake spans rather than other shortcuts with less generalization, and finally equips the model with the discrimination capacity between real and partially fake audios. Our submission ranked second in the partially fake audio detection track of ADD 2022.

Index Terms— Anti-spoofing, partially fake audio detection, audio deep synthesis detection challenge

1. INTRODUCTION

The past few years have witnessed significant advances in speech synthesis and voice conversion technologies, and recently emerged adversarial attacks, such that even humans may not be capable to distinguish the real users' speech from the synthesised speech [1–23]. Such technologies can undermine the robustness of broadly implemented biometric identification models, e.g. automatic speaker verification (ASV) models, and can be harnessed by in-the-wild attackers for criminal usage. For instance, an attacker can generate fake audios to manipulate the voiceprint-based security entrance system to accept the attacker falsely, and get access to normally protected information and valuables. Additionally, an imposter can call the bank center, fool the biometric identification system to accept him/her as a registered user, and transfer money to the imposter's account. Considering the severe harm caused by synthesized fake audio, it is critical to devise methods to tackle such threats.

The ASVspoof challenge [1–4], a community-led challenge, arouses the attention from both the industry and the academia to tackle the spoofing audios in both physical access and logical access. In logical access, attacks are mainly from synthesized audios by advanced speech synthesis and voice conversion models, while in physical access, replayed audios are adopted as attacks. The challenge attracts various international teams, and various high-performance anti-spoofing models have been proposed to address the two kinds of attacks mentioned above. The adversarial attacks for ASV and anti-spoofing models have been well investigated [6–9, 14, 15]. To solve further challenging attack situations in realistic applications, the first Audio Deep Synthesis Detection challenge (ADD 2022) [5] extends the attack scenarios to fake audio detection. They consider the fake audios perturbed by diverse background noise, and attacks from the latest speech synthesis and voice conversion models. Additionally, the organizers propose partially fake audio detection track, where the attacks are composed of hiding small fake clips into real speech. Partially fake audios are dangerous, and ADD 2022 is the first challenge attempting to tackle this type of brand new attacks, which is an open question, and is the focus of this paper

During generation of partially fake audio, only small clips of synthetic speech are inserted, and thus the fake audio contains a large proportion of genuine user's audio. Through experimentation, we find it is hard to distinguish the fake and real audios by directly implementing the previous state-of-the-art spoofing countermeasure models, such as Light Convolutional Neural Network (LCNN) [24] and Squeeze-and-Excitation Network (SENet) [25]. To allow the model discover the small anomalous clip in real speech, we design a proxy task to make the model answer "where are the start and end points" of such anomalous clips. During training, the anti-spoofing model not only has to predict the fake or real label for each utterance, but also to find the start and end positions of the fake clips within the utterance. Identifying the time segments of the fake clips is similar to *extraction-based question-answering*, which determines the answer span in a document. Also, to further improve the capacity of the anti-spoofing model to tackle the "question-answering" task, we introduce the self-attention [26] strategy. The experimental results illustrate the effective discrimination capacity of our proposed method between real and partially fake audios.

Our main contributions are two-fold:

- We proposed a brand new framework inspired by the extraction question-answering strategy for locating the fake regions

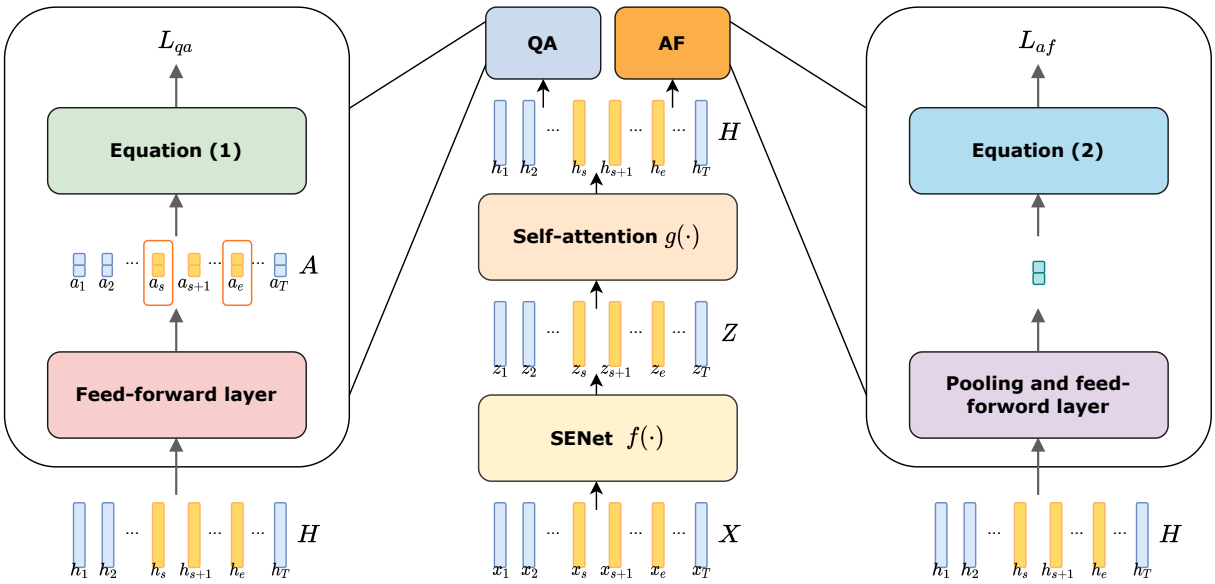

Fig. 1. The proposed framework. X, Z, H, A are the acoustic features, hidden features, bottleneck features and the output for Question-answer layer, respectively. f and g are the SENet feature extractor and self-attention layer, corresponding to (1)-(8) and (9) in Table 1, respectively. QA and AF are the question-answering (fake span discovery) and anti-spoofing layers with loss calculation procedures respectively.

in the fake, overall input audio, in order to improve the performance for partially fake audio detection.

- We further equipped the fake span discovery strategy with the self-attention mechanism to get a better detection capacity.

Also, our submission ranked second in the partially fake audio detection track of ADD 2022.

This paper is organized as follows: Section 2 introduces the proposed method, namely self-attention-based question-answering framework for partially fake audio detection. Section 3 presents experimental setups, followed by section 4 reporting on experimental results and analysis. Section 5 presents the conclusion.

2. METHODOLOGY

In this section, we will introduce the anti-spoofing method equipped with the proposed question-answering strategy and self-attention mechanism. We firstly present the details of the proposed framework. And then we will clarify the rationale of the proposed method.

2.1. Proposed anti-spoofing model

We adopt the base model SENet [25], which is a variant of ResNet [27] equipped with squeeze-and-excitation networks [28], and we perform some modifications to that model. The modified model architecture is shown in Table 1. Let $X = [x_1, x_2, ...x_T]$ denote the T frames of input acoustic features. The extracted hidden features by the SENet feature extractor are denoted as $f(X) = Z = [z_1, z_2, ...z_T]$, where f is the (1)-(8) layers in Table 1. The bottleneck features are denoted as $g(Z) = H = [h_1, h_2, ...h_T]$, where g is the self-attention layer, the layer (9) in Table 1, and $h_t \in R^n$. The self-attention layer is one layer of transformer [26]. The question-answering layer (a) is one fully-connected layer with the input dimension as n, the dimension for h_t, and with output dimension as 2. The 2 dimensions represent how likely will h_t be the start or end position of the fake clip. Given H as the input, the question-answering layer will output $A = [a_1, a_2, ...a_T]$, where $a_t \in R^2$. The question-answering loss L_{qa} is denoted as:

$$L_{qa} = -(log\frac{exp(a_s^1)}{\sum_{t=1}^{T} exp(a_t^1)} + log\frac{exp(a_e^2)}{\sum_{t=1}^{T} exp(a_t^2)}), \quad (1)$$

where s and e are the start and end positions for the fake clip, a_t^1 and a_t^2 are the values for first and second dimensions of a_t at the t^{th} frame. We will not incorporate the L_{qa} for training with real utterances. For the pooling layer (b), there are three pooling strategies in this paper, average pooling, self-attentive pooling (SAP) [29] and attentive statistics pooling (ASP) [30]. Based on the bottleneck features H, the pooling layer (b) followed by the prediction layer (c) will output $S = [s_0, s_1]$, indicating whether the utterance is fake or real. The anti-spoofing loss L_{af} is denoted as:

$$L_{af} = -log\frac{exp(s_l)}{\sum_{j=0}^{1} exp(s_j)}, \quad (2)$$

where $l \in \{0, 1\}$ is the target label. The final loss is

$$L = L_{qa} + L_{af}. \quad (3)$$

2.2. Rationale

In the partially fake audio detection track, there is only a small proportion of fake audio frames in the overall piece of input speech. Previous state-of-the-art anti-spoofing models [24, 25] tackle the problem of identifying whether a whole audio utterance is real or fake. Hence, previous strategies are not designed to identify anomalous regions within one utterance. Thus the previous models intuitively attain the ability to distinguish between utterances but there is no guarantee that such models can discover the abnormal regions within

Table 1. Proposed anti-spoofing model.

layer	Type	Filter / Stride	Output shape
(1)	Conv	$7 \times 7/1 \times 2$	$16 \times 501 \times 40$
(2)	BatchNorm	–	–
(3)	ReLU	–	–
(4)	MaxPool	$3 \times 3/1 \times 2$	$16 \times 501 \times 20$
(5)	SEResNet Module$\times 3$	–	$16 \times 501 \times 20$
(6)	SEResNet Module$\times 4$	–	$32 \times 501 \times 10$
(7)	SEResNet Module$\times 6$	–	$64 \times 501 \times 5$
(8)	SEResNet Module$\times 3$	–	$128 \times 501 \times 3$
(9)	Self-attention	–	501×384
(a)	Question-answering	–	501×2
(b)	Pooling	–	384
(c)	Prediction	–	2

a single utterance. To evaluate the performance of the previous state-of-the-art anti-spoofing models, we direct train binary classification anti-spoofing models for the partially fake audio detection task with reference to previous papers. We discover that these well-trained models do not have the discriminative ability for the adaptation set provided by the organizers of ADD 2022. A reasonable explanation is that the models may have learned some shortcuts to differentiate the audios with real and fake labels in the training set, but what the models have learned can not generalize to the adaptation set. In other words, the models cannot discover the fake regions for fake audio detection.

Thus, to regularize the model to learn to distinguish between the real and partially fake audios, we propose a proxy task to let the model discover the abnormal parts within a piece of partially fake audio. The proposed anti-spoofing model has to predict not only whether the input utterance is real or fake, but also output the start and end of each anomalous region. We name this proxy task as question-answering, or fake span discovery proxy task, in which the model has to answer "where is the fake clip" in a piece of partially fake audio. The extraction-based question-answering models in natural language processing (NLP) often take a question and a passage as input, build representations for the passage and the question respectively, match the question and passage embeddings, and output the start and end positions within the passage as the answer. We adopt the analogy of extraction-based question-answering here. The passage is the partially fake utterance, and the answer span is the time of the fake clip. By the question-answering proxy task, the model can learn to find the fake clips within an utterance, thus benefiting the model to distinguish between the audios with or without fake clips. Moreover, the self-attention module followed by the question-answering task addresses the model to attend on the fake regions, and helps reduce the question-answering loss, resulting in better discrimination capacity between real and partially fake audios.

3. EXPERIMENTAL SETUP

3.1. Data preparation

3.1.1. Dataset construction

The training set and dev set, which are based on Mandarin publicly available corpus AISHELL-3 [31], provided by the organizers of ADD 2022, cannot be directly adopted to tackle the problem of partially fake audio detection track, as the whole utterance sample in them is either real or fake. During the training phase, for constructing fake audios, we generate the partially fake audio by inserting a clip of audio into the real audios. The inserted clips are derived from three sources: 1). fake audios in the training and dev set provided by ADD 2022. 2). Real audios other than the victim audio in the training and dev set. 3). audios re-synthesised by the traditional vocoders, including Griffin-Lim [32] and WORLD [33], based on the real audios in the training and dev set. It is hard to train text-to-speech (TTS) or voice conversion (VC) models based on the limited real data provided by the organizer, so we choose the traditional vocoders, namely Griffin-Lim and WORLD, to increase the diversity of fake audios. As for the validation set, we adopt the adaptation set consisting of partially fake audios synthesised by ADD 2022 for selecting the models. We report the equal error rate (EER) for the testing set released by the organizer, as EER is the main evaluation metric for the partially fake audio detection track.

Table 2. The EERs with (w/) or without (w/o) self-attention.

FFT window size	w/o attention	w/ attention
384	23.6%	14.3%
768	22.0%	17.9%

3.1.2. Input representations

Mel-spectrograms, which are based on short-time Fourier transform (STFT) where the window size of fast Fourier transform (FFT) is varied from 384 to 768, the hop size is 128, and the number of output bins is 80, are used as input features for most of our experiments and are denoted by MSTFT in following sections. Besides spectral features, some extra experiments are operated on cepstral and NN-based features to increase diversity for achieving a better performance in the stage of fusion. The FFT window size, hop size, and number of output bins are fixed to 384, 128, and 80 respectively for Mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and SincNet [34], as we find the FFT window size of 384 performs well as shown in Table 3.

3.1.3. Data augmentation

We perform on-the-fly data augmentation by adding noise from MUSAN dataset [35], performing room impulse response (RIR) simulation [36] and applying codec algorithms (a-law and μ-law) [37].

3.2. Implementation details

The backbone model is shown in Table 1. Three kinds of attention, average pooling (Avg), attentive statistics pooling (ASP) and self-attentive pooling (SAP) are adopted for experiments. All the models are optimized by Adam with the learning rate of 0.001 and weight decay as $1e^{-4}$.

4. EXPERIMENTAL RESULTS AND ANALYSIS

First of all, we illustrate that the question-answering (QA) strategy drastically decreases the EERs. We conduct experiments with and without the QA strategy. The experimental results show that the trained models without the QA strategy attain the EERs of around

Table 3. The EERs using MSTFT features. w/o or w/ mean with or without. w/ or w/o re-synthesis correspond to using the re-synthesised audios by Griffin-Lim and WORLD or not.

feature	FFT window size	pooling method	w/o augmentation		w/ augmentation	
			w/o re-synthesis	w/ re-synthesis	w/o re-synthesis	w/ re-synthesis
MSTFT	384	Avg	14.3%	19.9%	11.9%	14.2%
	512	Avg	13.2%	20.5%	13.0%	14.8%
	640	Avg	18.5%	19.9%	18.9%	13.3%
	768	Avg	17.9%	16.8%	14.8%	12.6%
MSTFT	384	SAP	16.9%	17.5%	15.6%	12.6%
	512	SAP	17.0%	18.0%	13.9%	12.5%
	640	SAP	12.1%	15.3%	15.3%	11.1%
	768	SAP	15.2%	17.8%	11.7%	14.8%
MSTFT	384	ASP	17.3%	15.9%	14.9%	11.9%
	512	ASP	14.9%	15.8%	12.9%	11.1%
	640	ASP	17.5%	15.9%	15.8%	11.2%
	768	ASP	14.8%	17.9%	14.5%	22.1%

Table 4. The EERs for three different features

feature	MFCC	LFCC	SincNet
EER	12.5%	11.1%	16.1%

40%, which indicates that such models can not distinguish the partially fake audios from the genuine audios. Due to the poor performance of models without the QA strategy on the adaptation set, we decide not to submit the results on testing sets of such models to the leaderboard to save the submission times.

Next, we verify the effectiveness of the self-attention layer by Table 2. As the input and output feature dimensions after the self-attention layer are the same, the model without the self-attention layer can be constructed by directly removing (9) in Table 1. In the following experiments, the performances on the testing set will be directly displayed. We show the EERs under two settings of FFT window size due to space limitation, and the other settings are with the same trend. Table 2 shows that the improvements are significant in two settings with different window sizes. The EERs decrease 9.3% and 4.1% absolute after adding self-attention for the FFT window size of 384 and 768 respectively, which illustrates the significant improvements by introducing the self-attention layer.

Therefore, the model with self-attention will be adopted for the following experiments, unless specified otherwise. In the main experiments as shown in Table 3, the input representations are MSTFTs with hop size of 128, output bins as 80, and FFT window size ranging from 384-768. Table 3 exhausts the experimental settings under four different window sizes, three pooling strategies, whether to use the data augmentation and whether to use the re-synthesised fake audios by Griffin-Lim and WORLD. We have the following observations. First, EERs are improved with the help of data augmentation in most of the setups. Secondly, enlarging the training set by the re-synthesised data usually benefits the EERs when data augmentation is conducted. Lastly, the SAP and ASP pooling significantly improve the EERs when both data re-synthesis and augmentation are applied. We also can observe that the best EER for a single model is 11.1% shown in Table 3.

In order to increase diversity of models for achieving a better performance in the stage of model fusion, we further take MFCC, LFCC and SincNet as input features to train the models. We cannot exhaust all the settings due to limited computing resources, thus we refer to Table 3 to select the setting to conduct the experiments. We fix the FFT window size as 384, apply only ASP pooling, adopt data augmentation and the re-synthesised data. We observe from Table 4 that the LFCC feature gets EER as 11.1%, reaching the best single model performance in our experimental settings. For the further work, we plan to explore the potential of different front-end features to get better performance.

For the fusion method, we tried average fusion, weighted average fusion, min fusion and max fusion. The best submission, which is fused by the average scores of the top 5 models, achieves the best 7.9% EER and ranks second in partially fake audio detection track.

5. CONCLUSION

Inspired by extraction-based question answering, this paper proposes a self-attention-based, fake span discovery strategy. The proposed strategy tasks the anti-spoofing model to predict the start and end position of the fake clip within the partially fake audio, address the model's attention into discovering the fake spans rather than other patterns with less generalization, and finally equips the model with the discriminate capacity between real and partially fake audios. Our final submitted model gave 7.9% EER, and ranked 2nd in the partially fake audio detection track of ADD 2022. Such a strategy can be model-agnostic and feature-agnostic. Our future work will explore the potential of the proposed strategy by adopting other backbone anti-spoofing models and front-end features.

6. ACKNOWLEDGEMENTS

This research is funded by the Centre for Perceptual and Interactive Intelligence, an InnoCentre of The Chinese University of Hong Kong. This work was done while Haibin Wu was a visiting student at The Chinese University of Hong Kong. Haibin Wu is supported by Google PHD Fellowship Scholarship.

7. REFERENCES

[1] Z. Wu, T. Kinnunen, et al., "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] T. Kinnunen, M. Sahidullah, et al., "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[3] M. Todisco, X. Wang, et al., "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[4] J. Yamagishi, X. Wang, et al., "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP*. IEEE, 2022.

[6] H. Wu, A. T. Liu, and H.-y. Lee, "Defense for black-box attacks on anti-spoofing models by self-supervised learning," *arXiv preprint arXiv:2006.03214*, 2020.

[7] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense against adversarial attacks on spoofing countermeasures of asv," in *ICASSP*. IEEE, 2020, pp. 6564–6568.

[8] Z. Peng, X. Li, and T. Lee, "Pairing weak with strong: Twin models for defending against adversarial attack on speaker verification," in *Proc. INTERSPEECH*, 2021.

[9] H. Wu et al., "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *ICASSP*. IEEE, 2021, pp. 6718–6722.

[10] S. Liu, H. Wu, H.-y. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *ASRU*. IEEE, 2019, pp. 312–319.

[11] X. Li et al., "Replay and synthetic speech detection with res2net architecture," in *ICASSP*. IEEE, 2021, pp. 6354–6358.

[12] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," *arXiv preprint arXiv:2107.08803*, 2021.

[13] H. Wu, Y. Zhang, Z. Wu, D. Wang, and H.-y. Lee, "Voting for the right answer: Adversarial defense for speaker verification," *arXiv preprint arXiv:2106.07868*, 2021.

[14] H. Wu et al., "Improving the adversarial robustness for speaker verification by self-supervised learning," *arXiv preprint arXiv:2106.00273*, 2021.

[15] H. Wu et al., "Spotting adversarial samples for speaker verification by neural vocoders," *arXiv preprint arXiv:2107.00309*, 2021.

[16] 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.

[17] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *APSIPA*. IEEE, 2014, pp. 1–5.

[18] M. R. Kamble et al., "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[19] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *ICASSP*. IEEE, 2020, pp. 6589–6593.

[20] W. Chenglong, Y. Jiangyan, et al., "Global and temporal-frequency attention based network in audio deepfake detection," *Journal of Computer Research and Development*, 2021.

[21] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection.," in *INTERSPEECH*, 2021, pp. 886–890.

[22] J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," in *INTERSPEECH*, 2021, pp. 1654–1658.

[23] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[24] 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.

[25] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *arXiv preprint arXiv:1904.01120*, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[28] 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.

[29] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Interspeech*, 2017, pp. 1517–1521.

[30] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018, pp. 2252–2256.

[31] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[32] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, pp. 236–243, 1984.

[33] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[35] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[36] T. Ko, V. Peddinti, et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017, pp. 5220–5224.

[37] C. Recommendation, "Pulse code modulation (pcm) of voice frequencies," in *ITU*. 1988.