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

Attention-based Convolutional Neural Network for ASV Spoofing Detection

Hefei Ling, Leichao Huang, Junrui Huang, Baiyan Zhang and Ping Li

School of Computer Science and Technology, HUST, Wuhan, China
{lhefei,lchuang,junrui_h,zhangbyxy}@hust.edu.cn,lpshome@mail.hust.edu.cn

Abstract

In recent years, automatic speaker verification (ASV) algorithms have undergone significant progress. They have been widely deployed in different applications, but the ASV systems are vulnerable to spoofing attacks, such as impersonation, replay, text-to-speech, voice conversion and the recently emerged adversarial attacks. To improve the robustness of the ASV system, researchers have designed anti-spoofing systems to resist spoofing attacks. While previously proposed systems have shown to be effective for spoof attacks detection, they are all ensemble methods based on different speech representations and architectures at the cost of increased model complexity, with similar performance not being achieved with single systems. This paper proposes an attention-based single convolutional neural network to learn discriminative feature embedding for spoof detection, achieving performance comparable to ensemble methods. The key idea is to decrease the information redundancy among channels and focus on the most informative sub-bands of speech representations. The experiments show that our proposed single system achieves an equal error rate of 1.87% on the evaluation set of ASVspoof 2019 Challenge, outperforming all single systems and comparable to the second-ranked system (EER 1.86%) among all known systems.
Index Terms: anti-spoofing, automatic speaker verification, attention-based cnn

1. Introduction

Automatic Speaker Verification(ASV) systems aim at confirming a claimed speaker identity by a spoken utterance, and that is used for a wide range of application services in recent years [1–3]. However, ASV systems are vulnerable to spoofing attacks, such as impersonation (mimics or twins), replay (pre-recorded audio), text-to-speech(TTS), voice conversion (VC), and the recently emerged adversarial attacks [4, 5]. The recent advances in VC and TTS technologies have produced high-quality natural-sounding speech [6], and have shown a potential threat to ASV systems [7, 8].

To solve the problem of spoofing attacks, the ASVspoof challenge series [9–11] has been held in recent years, which provides datasets and metrics for anti-spoofing speaker verification research. The previous ASVspoof Challenges focused on raising awareness and fostering solutions to address spoofing attacks generated by replay, TTS and VC. Especially, ASVspoof 2019 contains all previous attacks, which divides the attacks into logical access (LA) and physical access (PA). The LA task focuses on detecting TTS and VC attacks, and the PA task focuses on the detection of replay attacks. In this study, we focus on the LA attacks.

Most previous methods have focused on the design of specific front-ends, and results from the bi-annual ASVspoof evaluations also show that effective countermeasures demand handcrafted features specially designed such as Zero Time Windowing Cepstral Coefficients (ZTWCC), Linear Frequency Cepstral Coefficients (LFCC) and Constant-Q Cepstral Coefficients (CQCC) [12]. Given the diversity in spoofing attacks, in order to further improve the performance, some researchers introduced model fusion based on different features [13–15] at the cost of increased model complexity. It shows that no single front-end can detect reliably the full range of artefacts produced by different spoofing attacks. Moreover, some researches [16–18] show that not all frequency bands are helpful for these spoofing tasks and that these can only be detected reliably using front-ends that have high spectral resolutions in the same bands [19, 20]. However, conventional cepstrum processing only treats them in the same way, instead of emphasizing information at the sub-band level, which may affect anti-spoofing performance. Besides, non-local correlations exist in a T-F spectrogram along the frequency axis. A typical example is the correlations among harmonics, which has been shown to be helpful for spoofing detection, but simply stacking several 2D convolution layers with small kernels cannot capture such global correlation.

To address these problems, this paper proposes an attention-based system for spoofing detection. Attention can be viewed as a tool to learn the most informative components of an input signal, which focuses on essential features and suppressing unnecessary ones. Attention-based methods have been applied in classification and recognition tasks, e.g. language models [21], face recognition [22], image classification [23] and ASV replay attack detection [24, 25]. Precisely, the proposed attention module consists of two blocks named frequency attention block (FAB) and channel attention block (CAB). The frequency attention block aims to learn non-local correlations in a T-F spectrogram along the frequency axis and focus on the essential sub-bands, and the channel attention block aims to learn the inter-channel relationships to decrease the information redundancy among channels. Experiments show that our proposed method outperforms all existing single systems on the ASVspoof 2019 LA dataset and ranks between the second and third places among all participating systems.

The rest of this paper is organized as follows. Section 2 describes the general framework of our proposed systems and then introduces the two attention blocks in detail. Section 3 shows the experimental setup and results. Finally, we summarize the conclusions derived from this research in Section 4.

2. Our proposed methods

This section will first present a general framework of our proposed system and then introduce the two attention blocks in detail. Finally, we describe how to aggregate them together for ASV spoofing detection.

http://dx.doi.org/10.21437/Interspeech.2021-1404

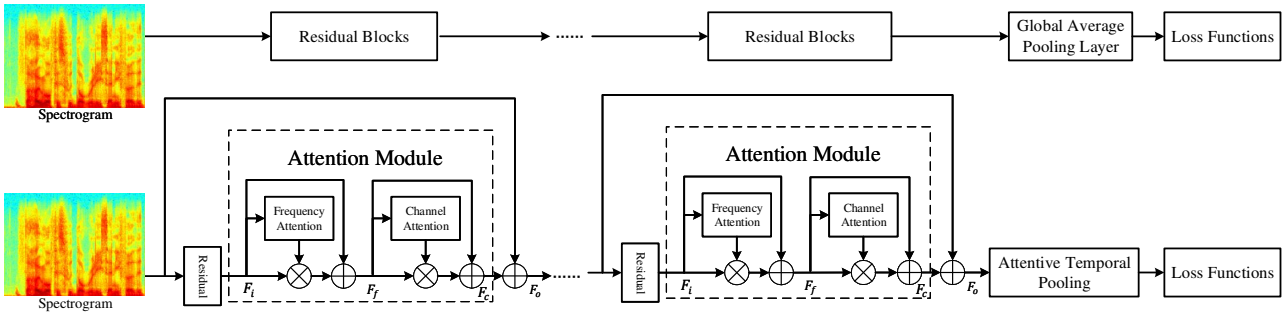

Figure 1: *Schematic diagram of our proposed architecture. Top of this figure is a standard process for ResNet and the bottom is ours. Given an intermediate feature map F_i, the attention module first generates a Frequency refined feature F_f, then yields a channel refined feature F_c. $\otimes$ means matrix multiplication and $\oplus$ means element-wise summation.*

2.1. Overview architecture

As illustrated in Figure 1, We adopt the network architecture adapted from [26]. Building upon the original CNN architectures, we add the attention module to the top of each residual bottleneck of the ResNet structure. The proposed attention module consists of two blocks, frequency attention block and channel attention block, to capture frequency relationship and channel relationships. Specifically, given an input feature map F_i, the attention module first generates an inter-frequency relationship matrix and the weighted feature by matrix multiplication sequentially, and then yields the frequency-refined feature F_f by element-wise summation. Inspired by Resnet, this residual learning can help improve the stability of the training process. Similarly, we can then obtain the channel-refined feature F_c in the same way. In addition, considering that global average pooling may reduce detection performance, we employ attentive temporal pooling to assign higher importance to particular segments of the input. As a result, our proposed system can reduce the redundancy of information and focus on the most crucial part through the methods mentioned above, thus obtaining better performance.

2.2. Frequency attention block

The prior studies (Section 1) revealed that not all frequency bands are helpful for these spoofing tasks, which inevitably results in information redundancy and even makes the feature has a risk of over-fitting. Besides, non-local correlations, such as harmonics, exist in a time-frequency spectrogram along the frequency axis. However, simply stacking several 2D convolution layers with small kernels cannot capture such global correlation. Therefore, we design the frequency attention block to be inserted at the top of each residual block to enable the model to have a full-frequency receptive field and focus on the most informative frequency bands. Figure 2 shows the details of the proposed frequency attention block.

With the input feature $F_i \in R^{C\times F\times T}$, the FAB first squeezes the feature map along the channel axis and time axis using both global average pooling and max pooling parallelly, which generates two feature vectors F_f^{avg} and F_f^{max}. Then those two features are concatenated to form the aggregated feature F_f^{cat}, followed by 2D 1x1 convolution to obtain an intermediate feature map F_f^{pool}. It can be formulated as:

$$F_f^{pool} = Conv(F_f^{avg} \copyright F_f^{max}) \tag{1}$$

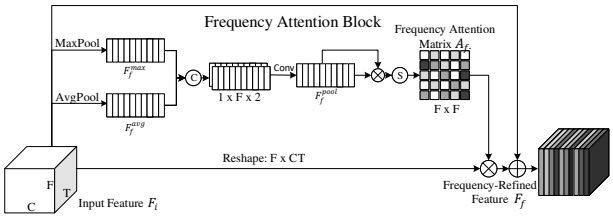

Figure 2: *Frequency Attention Block (FAB). In this figure, © denotes concatenation operator, $\otimes$ denotes matrix multiplication, Ⓢ means softmax activation and $\oplus$ means element-wise summation. And softmax operation is performed on each row of the matrix*

where $F_f^{pool} \in R^{1\times F\times 1}$, © means concatenate two features together. After that, we first applied reshape and calculate its autocorrelation matrix. Then softmax operation is performed on each row of the matrix to get the frequency attention matrix A_f.

$$A_f = Softmax(F_f^{pool} \otimes (F_f^{pool})^{\mathrm{T}}) \tag{2}$$

Finally, the input feature F_i is matrix multiplied by the frequency attention A_f, and then obtained a frequency-refined feature $F_f \in R^{C\times F\times T}$ via residual shortcut learning.

$$F_f = F_i \oplus (\alpha \times (A_f \otimes F_i)) \tag{3}$$

where α is a learnable parameter with initialization of 0 to decrease the difficulty of convergence process in first few training epochs.

2.3. Channel attention block

For a standard convolutional network, as the convolutional layer deepens, the number of channels increases, which inevitably results in an information redundancy among channels. To further improve the system's performance, the attention module also contains a channel attention block to learn the inter-channel relationships of an intermediate feature map. As Figure 3 illustrated, following the same strategy used in frequency attention block, we first apply average-pooling and max-pooling along the frequency axis and time axis,and then sum them together to form an efficient feature descriptor, followed by convolution operation, and softmax layer is conducted to obtain the final channel attention matrix. It can be formulated as:

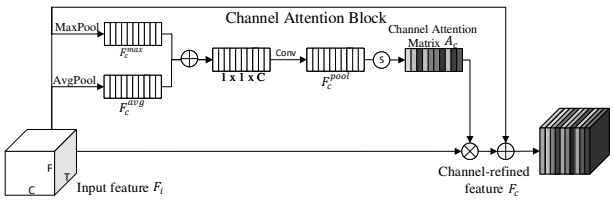

Figure 3: *Channel Attention Block (CAB). In this figure, $\oplus$ denotes element-wise summation, $\otimes$ means matrix multiplication and Ⓢ denotes softmax activation.*

Table 1: *Summary of the ASVspoof 2019 LA dataset*

	bonafide	**Spoof**	
	utterance	utterance	attacks
Training	2580	22800	A01-A06
Development	2548	22296	A01-A06
Evaluation	7533	63882	A07-A19

$$A_c = Softmax(Conv(F_c^{avg} \oplus F_c^{max})) \quad (4)$$

Finally, the input feature F_i is matrix multiplied by the channel attention A_c and then obtained a channel-refined feature $F_c \in R^{C\times F\times T}$ via residual shortcut learning.

$$F_f = F_i \oplus (\beta \times (A_c \otimes F_i)) \quad (5)$$

where β is a learnable parameter with initialization of 0 to decrease the difficulty of convergence process in first few training epochs.

2.4. Attention block integration designs

The two attention blocks can be individually or combined as a module integrated into the existing network. In this subsection, we introduce three different combined designs: (1) Sequential, in which the two blocks are combined in a sequential way; (2) Seq-inversed, in which the channel attention block is moved before the frequency attention block and (3) Parallel, in which the two blocks are combined in a parallel way. The details are illustrated in Figure 4.

3. Experiments

3.1. Datasets and Evaluation metrics

The ASVspoof 2019 challenge provides a standard database [27] for anti-spoofing, which contains two subset evaluations: physical access (PA) and logical access (LA). All our experiments are conducted under the LA subset. As Table 1 described, the LA subset is partitioned into three part for training, development and evaluation, and each part includes genuine speech and different kinds of TTS and VC spoofing attacks. Training and development sets share the same 6 attacks (A01-A06), consisting of 4 TTS and 2 VC algorithms. There are 13 attacks (7 TTS and 6 VC) in the evaluation set. It is noted that the evaluation set includes only two known attacks (A16, A19) and 11 unknown attacks (A07-A15, A17, A18).

We evaluate our model performance with the equal error rate (EER) and the tandem detection cost function (t-DCF) metrics. The t-DCF takes both the ASV system and spoofing countermeasure errors into consideration. With a fixed ASV system,

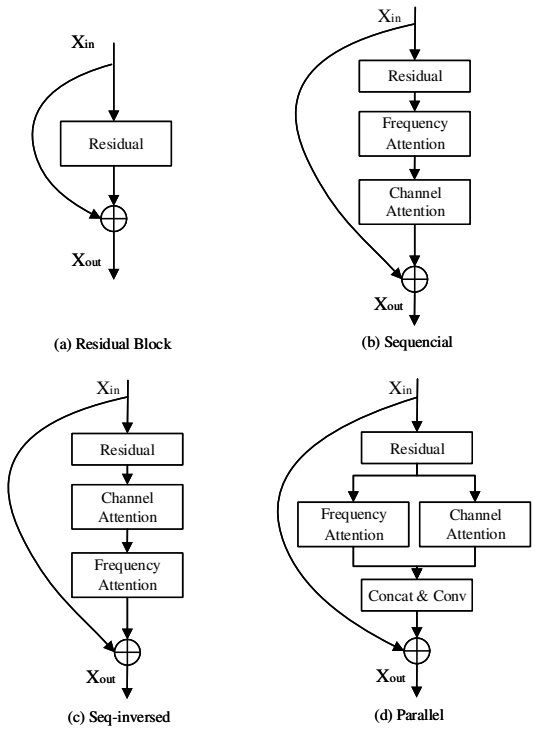

Figure 4: *Attention block integration designs. The two attention block are aggregated in three ways: (b)sequentially, (c)Seq-inversed and (d)parallel.*

the lower the t-DCF, the better the anti-spoof system performance. Further details of t-DCF can be found in [28].

3.2. Details of systems implementation

In this study, we extract the log power magnitude spectrogram of the utterances as input to the proposed system. The spectrograms are extracted with a "Hanning" window having the size of 25 ms and step-size of 10 ms, and 512 FFT points are applied. To form a consistent input representation, we set 750 frames as the fixed length and replicate the spectrogram features along the time axis if the duration is shorter than 7.5 seconds, and we randomly choose a consecutive piece of frames and discard the rest for long utterance. Therefore, the original input spectrogram has a shape of 750 × 257, where 750 is the number of frames and 257 is the number of FFT bins. None additionally preprocessing techniques such as speech activity detection or dereverberation was explored in these systems.

We use the network architecture from [26] with Softmax loss function as our backbone to train the baseline model. The architecture is based on deep residual network ResNet18 [29], where the global average pooling layer is replaced by an attentive temporal pooling layer. We use 256 for the embedding dimension. All our systems are trained with Softmax loss, except for system Sequential† constrained with OC-Softmax. The hyper-parameters of the OC-Softmax loss function is the same as in [30].

We implement the algorithms in the PyTorch framework [31]. We use Adam optimizer with the $\beta 1$ parameter set to 0.9 and the $\beta 2$ parameter set to 0.999 to update the weights in the model. The batch size is set to 32. The learning rate is initially set to 0.0003 with half decay for every 10 epochs. We trained

Table 2: *Results on the development set and the evaluation set of the ASVspoof 2019 LA scenario. † means the system constrained with OC-Softmax loss.*

Model	Dev Set		Eval Set	
	EER(%)	t-DCF	EER(%)	t-DCF
Baseline	0.44	0.014	3.26	0.081
Frequency	0.39	0.012	2.70	0.069
Channel	0.38	0.012	2.81	0.076
Sequential	0.34	0.010	2.38	0.062
Seq-inversed	0.43	0.012	2.55	0.071
Parallel	0.43	0.012	2.53	0.067
Sequential†	0.16	0.004	1.87	0.051

the network for 100 epochs, then we select the model with the lowest validation EER and report the results evaluate on evaluation set. For the system with Softmax loss, the anti-spoof system's output is directly adopted as the countermeasure (CM) score; for the system with OC-Softmax loss, the final CM score is the cosine similarity between the speech embedding and the weight vector in OC-Softmax.

3.3. Results and discussion

3.3.1. Evaluation of attention block

To evaluate the performance of our proposed method, we first conduct comparative experiments for each attention block separately and then analyze three different ways of aggregating the proposed frequency attention block and channel attention block. Finally, we train the proposed system with other loss functions, such as OC-Softmax, which introduced in [30] and demonstrated an elegant way to against unknown spoofing attacks. The results of the system aggregated with different attention block are presented in Table 2 on both the development and evaluation sets. From the table, we can observe that whether aggregated with frequency attention block or channel attention block, both evaluate results achieve better performance than baseline, confirming that the proposed attention module can capture more global channel information and focus on the most crucial part. While aggregated with both attention blocks, the sequential order performs slightly better than the other two ways. Besides, we experimentally verify that the proposed attention module can be combined with the existing state-of-the-art algorithms for a better spoof detection performance.

3.3.2. Comparison with other systems

We also compared the proposed system with various systems on the evaluation set of ASVspoof 2019 logical access corpus. We consider not only some of the top-performing systems of the ASVspoof 2019 challenge but also recently published works. Those systems contain the two ASVspoof 2019 baseline systems B1 and B2, and the top four performance systems in ASVspoof 2019 challenge results [11]. The latter are signified by their anonymous ASVspoof 2019 identifiers T05, T45, T60 and T24. In addition, recent published novel methods such as feature genuinization [34], data augmentation [35], novel loss function [30], and sub-bands model fusion [33] are also considered. For all methods with a reference, we obtained their results from their papers.

Table 3 reports the performance comparison of the proposed system to some of the single systems discussed above. It

Table 3: *Performance comparison with other single systems on the evaluation set of the ASVspoof 2019 LA scenario.*

System	EER (%)	min t-DCF
CQCC+GMM(B1)	9.57	0.237
LFCC+GMM(B2)	8.09	0.212
Wu et al. [32]	4.07	0.102
Tak et al. [33]	3.50	0.090
Chen et al. [34]	3.49	0.092
Rohan et al. [35]	3.13	0.094
Proposed	2.38	0.062
Zhang et al.† [30]	2.19	0.059
Proposed†	1.87	0.051

Table 4: *Performance comparison with other ensemble systems on the evaluation set of the ASVspoof 2019 LA scenario.*

System	EER (%)	min t-DCF
T24	3.45	0.095
Tak et al. [33]	2.92	0.074
T60	2.64	0.075
Proposed†	1.87	0.051
T45	1.86	0.051
T05	0.22	0.179

can be seen that our proposed system outperforms all other single systems (no model fusion) results in terms of both the performance metrics t-DCF and EER. Moreover, the proposed systems are comparable to the second-ranked system on the leader board of the ASVspoof 2019 Challenge for LA scenario, even though these competing systems are based upon an ensemble of comparatively complex neural network based architectures. Further detail presented in Table 4.

4. Conclusions

In this work, we propose an attention-based convolutional neural network to learn the global relationships of speech feature. Precisely, the proposed attention module consists of two blocks named frequency attention block and channel attention block. The frequency attention block aims to focus on the most informative sub-bands of Spectrograms, and the channel attention block aims to decrease the information redundancy among channels. We experimentally confirmed that our proposed attention module could work as an auxiliary tool to integrate with existing popular network architectures and loss functions to boost the spoof detection system's performance. The experiments also show that the proposed system (EER 1.87%) outperforms all existing single systems of the ASVspoof 2019 Challenge LA scenario and is comparable to the second system (EER 1.86%) among all participating systems. The future work will focus on extending the studies to replay attack detection and other multimedia forgeries.

5. Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61972169, in part by the National key research and development program of China(2019QY(Y)0202), in part by the Research Programme on Applied Fundamentals and Frontier Technologies of Wuhan(2020010601012182).

6. References

[1] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *IEEE signal processing society speech and language technical committee newsletter*, 2013.

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

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

[4] 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.

[5] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on gmm i-vector based speaker verification systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6579–6583.

[6] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[7] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," *arXiv preprint arXiv:1803.00860*, 2018.

[8] R. K. Das, T. Kinnunen, W.-C. Huang, Z. Ling, J. Yamagishi, Y. Zhao, X. Tian, and T. Toda, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," *arXiv preprint arXiv:2009.03554*, 2020.

[9] 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.

[10] 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.

[11] 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," *arXiv preprint arXiv:1904.05441*, 2019.

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

[13] 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.

[14] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[15] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," *arXiv preprint arXiv:1904.04589*, 2019.

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

[17] 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.

[18] 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.

[19] J.-w. Jung, H.-j. Shim, H.-S. Heo, and H.-J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoof 2019 challenge," *arXiv preprint arXiv:1904.10134*, 2019.

[20] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "An explainability study of the constant q cepstral coefficient spoofing countermeasure for automatic speaker verification," *arXiv preprint arXiv:2004.06422*, 2020.

[21] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[22] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 5595–5616, 2020.

[23] 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.

[24] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *Interspeech*, 2018, pp. 681–685.

[25] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6316–6320.

[26] 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.

[27] 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.

[28] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *arXiv preprint arXiv:1804.09618*, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards generalized voice spoofing detection," *arXiv preprint arXiv:2010.13995*, 2020.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[32] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," *arXiv preprint arXiv:2009.09637*, 2020.

[33] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," *arXiv preprint arXiv:2005.10393*, 2020.

[34] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop, Tokyo, Japan*, 2020, pp. 1–5.

[35] R. K. Das, J. Yang, and H. Li, "Data augmentation with signal companding for detection of logical access attacks," *arXiv preprint arXiv:2102.06332*, 2021.