INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Synthetic Voice Spoofing Detection based on Feature Pyramid Conformer

*Jingran Gong, Ning Chen**

East China University of Science and Technology, ShangHai, China
y80210024@mail.ecust.edu.cn, chenning_750210@163.com

Abstract

In speech anti-spoofing, artefacts used to detect spoofed speech are often located in specific sub-bands. Previous works often use Convolution Neural Networks (CNNs) as backbone which are good at capturing local features. However, if artefacts simultaneously exist in different sub-bands, CNNs cannot model this kind of information. Thus, we propose to use Feature Pyramid Conformer to solve this issue. Conformer can capture both local and global features. We aggregate the outputs of each Conformer block with Feature Pyramid Module. Through addition and lateral connection, the aggregation can be better integrated. Besides, to improve generalization of detecting unknown attacks, we propose to adopt Elastic penalty Margin Softmax. It can enhance intra-class compactness and inter-class discrepancy flexibly. Without data augmentaion, our system achieve an Equal Error Rate (EER) of 1.65% on the evaluation set of ASVspooof 2019 logical access, outperforming most existing systems.

Index Terms: Conformer, feature pyramid, anti-spoofing, speaker verification, loss function

1. Introduction

Automatic Speaker Verification aims at confirming whether a given speech signal proceeds from a given individual, and has broad application prospects in finance, banking, e-commerce and other fields[1]. However, it is possible to illegally pass ASV systems by imitation, voice conversion (VC), text-to-speech (TTS), replay and adversarial attacks, which brings severe challenges to the security of biometric verification systems. In recent years, speech synthetic and voice conversion technology have made significant progress in timbre and naturalness, which makes anti-spoofing face greater challenges.

In order to solve these security problems, the ASVspoof holds a competition every other year, which aims to encourage the development of effective countermeasures against unseen spoofing attacks in ASV systems[2, 3, 4]. Since ASVspoof2019, the challenge has been divided into two tracks: logical access (LA) and physical access (PA) scenarios. The former aims at detecting synthetic speech and the latter aims at detecting replay attacks. Many scholars have carried out multi-angle researches on synthetic speech detection and achieved great results. In out study, we focus on the LA attacks.

Convolutional neural network(CNN) approaches have been applied extensively in speech anti-spoofing such as ResNet[5], LCNN[6, 7, 8, 9], Res2Net[10]. CNNs are good at capturing local information progressively through a local receptive field layer by layer. On the contrary, they have difficulty in capturing long-range global representations. In recent years, Transformer[11] is very popular due to its efficiency in training and the ability of capturing long distance interaction, which is caused by self-attention mechanism. Although Transformers have achieved great performance in the field of ASR, they are difficult to extract fine-grained local feature templates and also require complex pre-training procedures.

Artefacts refer to signatures in spoofed speech that left behind by voice conversion and speech synthesis algorithms. It is well known that artefacts of spoofing attacks often reside in specific sub-bands or temporal segments [12, 13, 14, 15, 16], which indicates why CNNs can achieve good performance. However, if artefacts present in different filters or sub-bands at the same time, CNNs can not model such information. In [17], it has showed the merit of graph attention networks (GATs) to learn the relationships between cues in different sub-bands or temporal intervals by using self-attention mechanism. Therefore, we propose to use Conformer[18] as a powerful modeling tool to solve the problem. Conformer, a combination of CNN and Transformer, may exploit local and global cues with more distinguishable power.

Besides, previous studies[19, 20] indicate that low-level feature maps also assist in speech embedding extraction. Due to the depth of the network, there is a large semantic gap between the lower layer and the higher layer. The feature maps from higher layer are more discriminative. In order to aggregate both lower and higher layer feature maps and also enhance the discriminability of lower-layer feature maps, we propose to use Feature Pyramid Module (FPM). Through the top-down architecture with lateral connection, the module can produce a feature representation in which all levels are sematically strong.

One of the most important problems in anti-spoofing is the generalization to unseen spoofing attacks in the test stage[21]. In recent years, Softmax and AM-softmax are commonly used in anti-spoofing. There are also some studies focusing on designing more efficient loss functions. Chen[22] uses Large Margin Cosine Loss (LMCL) which reforms softmax loss as a cosine loss. LMCL can force DNN to learn the feature representation that can maximize inter-class variance and minimize intra-class variance. Zhang[23] believes that the distribution of various types of forgery attacks is not similar and proposed one-class softmax (OCsoftmax). However, these methods all proposed to incorporate a fixed penalty margin on loss function. Such learning objectives are unrealistic for data with different speakers and attacks, which may limit the discrimination and flexibility of anti-spoofing models. In our paper, we proposed to relax the fixed margin by Elastic Margin Softmax (EM-Softmax)[24] that allows flexibility for classification. The main idea is to extract random penalty margin values from a normal distribution in each training iteration. This makes the decision

*Corresponding author

10.21437/Interspeech.2023-1621

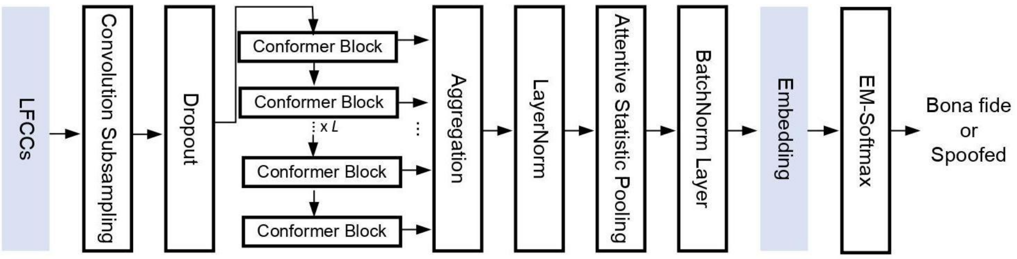

Figure 1: *The structure of our proposed methods*

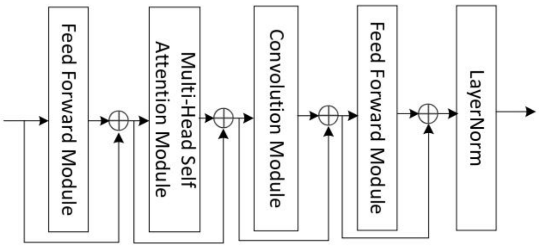

Figure 2: *The structure of the Conformer*

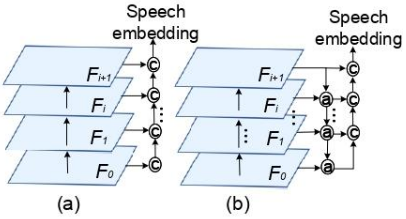

Figure 3: *Feature pyramid module. (a):aggregation w/o FPM. (b):aggregation with FPM.* ⓐ *denotes adding.* ⓒ *denotes concatenation.*

boundary more flexible and leaves space for class separability learning.

The rest of this paper is organized as follows. Section2 describes the general framework of our proposed systems and introduces the Feature Pyramid Conformer and Elastic Margin Softmax in detail. Section 3 shows the experimental setup. Section4 reports the result of our experiments. Finally, we summarize the conclusions derived from this research in Section 5.

2. Methodology

In this section, we will introduce the proposed Feature Pyramid Conformer (FP-Conformer) and Elastic Margin Softmax (EM-Softmax). The overall architecture is shown in Figure 2.

2.1. FP-Conformer

Firstly, the hand-craft feature LFCCs will pass through a convolution subsampling layer and dropout to reduce computational cost and prevent overfitting. Combined with CNN and Transformer, Conformer can capture both long-range global context dependencies and local details which may be helpful in learning discriminative cues between spoofing and bona fide speech. Conformer is composed with two half-step feed forward networks (FFNs), one Multi-head self-attention (MHSA) module and one convolution module, and each whole block is followed by LayerNorm. Each module adopts residule unit. The two half-step FNNs makes the Conformer block look like a Macaron and performs better than one single FNN[25]. From the Figure 2, we can see that convolution and self-attention are concatenated to achieve enhanced effect.

In anti-spoofing, it is important to learn the cues in the spoofed audio. Conformers are feed-forward architectures and use repeated Conformer blocks. Because of the depth in networks, there is large semantic gap between low-level and high-level feature maps. In order to aggregate all the outputs of each Conformer block, we adopt feature pyramid module (FPM) rather than concatenating the outputs directly as in[26]. There are L Conformer blocks in our architecture. As showed in Figure 3, for the output feature map F_i from i-th Conformer block, it will be added with the output of last block $(i+1)$-th from top to down. This strengthens the deep discriminative information and supplements the shallow complementary information.

2.2. Elastic Angular Margin Penalty-based Loss

An efficient loss function is also important in detection. The most common loss function is Softmax. It is defined as the combination of last fully connected layer, softmax function and cross-entropy loss, which can be formulated as follows:

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^T\boldsymbol{x}_i}} \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \{0,1\}$ are the embedding vector and label of the i-th sample. $\boldsymbol{w}_0$, $\boldsymbol{w}_1 \in \mathbb{R}^D$ are the weight vectors of two different classes, and N is the number of samples in a mini-batch.

Based on Softmax, the Addictive Margin (AM) Softmax loss function was proposed to replace the inner product operation of Softmax function with the cosine similarity operation in order to widen the inter-class margin in the embedding space that enhance the feature discrimination. It can be expressed as:

$$\mathcal{L}_{AMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\cos(\theta_{y_i})-m)}}{e^{\alpha(\cos(\theta_{y_i})-m)}+e^{\alpha\cos(\theta_{1-y_i})}} \tag{2}$$

$$\hat{\boldsymbol{w}}_j = \frac{\boldsymbol{w}_j}{||\boldsymbol{w}_j||} \tag{3}$$

$$\hat{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i}{||\boldsymbol{x}_i||} \tag{4}$$

$$\cos(\theta_{y_j}) = \hat{\boldsymbol{w}}_j^T\hat{\boldsymbol{x}}_i \tag{5}$$

where m is an additional penalty margin and α is a scaling factor for stabilizing training. θ_{y_i} is the angle between the weight $\boldsymbol{w}_{y_i}$ and the feature representation $\boldsymbol{x}_i$.

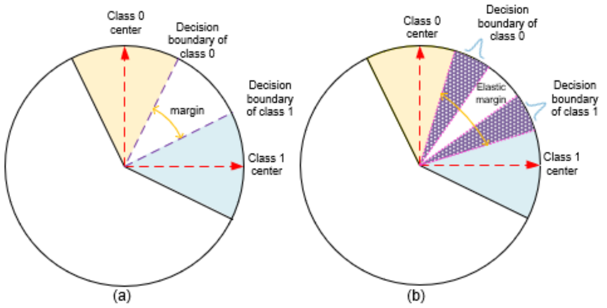

Figure 4: *The difference between AM-Softmax and EM-Softmax. (a):AM-Softmax. (b):EM-Softmax.*

Unlike the AM-Softmax that utilize a fixed margin value, we propose to adopt random margin penalty-based loss to anti-spoofing problem by randomly extracting margin values from a Gaussian distribution. In each iteration, the margin is different for each sample and changes in the next iteration. The Gaussian distribution can be expressed as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

where μ is the mean of the distribution and σ is the standard deviation. The EM-Softmax can be formulated as:

$$\mathcal{L}_{ES} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\alpha(\cos(\theta_{y_i})-E(m,\sigma)}}{e^{\alpha(\cos(\theta_{y_i})-E(m,\sigma))} + e^{\alpha(\cos(\theta_{1-y_i}))}} \tag{7}$$

where $E(m, \sigma)$ is a normal function that returns a random margin from (6) with the mean m and the standard deviation σ. The decision boundaries of AM-Softmax and EM-Softmax are illustrated in Figure 4. As we can see from the figure, rather than setting a fixed penalty margin value to train the model, the randomized margin penalty may give the model flexibility and generalization in detecting various and unseen synthetic speech. In anti-spoofing, there are many kinds of spoofing attacks. Gernaralization and flexibility are two important points. Considering multi speakers and spoofing attacks, it is reasonable to set different margins in loss function.

3. Experiments

3.1. Datasets

We only use the ASVspoof 2019 LA dataset[27] in our experiments. This dataset contains 17 different attacks. The LA dataset is divided into three subsets for training, development and evaluation. The training and development sets contain the same 6 attacks(A01-A06). The evaluation set contains 11 unseen attacks and 2 known attacks. The details of the LA dataset are shown in Table 1.

Table 1: *Summary of the ASVspoof2019 LA dataset*

Datasets	Bona fide	Spoofed	
	utterance	utterance	attacks
Training	2580	22800	A01-A06
Development	2548	22296	A01-A06
Evaluation	7355	63882	A07-A19

3.2. Evaluation Metrics

We use equal error rate (EER) and the minimum tandem detection cost function (min t-DCF) as the metrics for all experiments. For anti-spoofing task, EER is the value when the false rejection rate (FRR) and false accepytance rate (FAR) are equal. EER can reflect the security and accuracy of the system at the same time, and it is an important indicator to measure the performance of the biometric system. The min t-DCF[28] shows the impact of spoofing and the spoofing detection system upon the performance of an automatic speaker verification system.

3.3. Details of system implementation

In our experiment, we extract 60-dimensional linear frequency cepstral coefficients (LFCCs) from the utterances with MATLAB provided by the ASVspoof 2019 Challenge organizers. We set the frame size as 20ms and the hop size as 10ms. We use Pytorch framework to implement the FP-Conformer and EM-Softmax. The architecture takes the extracted LFCCs as input and outputs 256-dimentional embedding. We use the network architecture from[26] and improve it with FPM. We set the convolutional subsampling rate as 1/4. In the Conformer block, the encoder dimension is 256 and the number of attention heads is 4 for multi-headed self-attention; for convolution module, the kernel size is 15; for feed forward module, the linear hidden units is 2048. For the hyper-parameters in the loss functions, we set α=20, m=0.9 for all the loss functions in our experiments and set σ=0.0125 for EM-Softmax. 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 FP-Conformer. The batch size is set to 64. The learning rate is initially set to 0.0003 with half decay for every 10 epochs. We trained the network for 100 epochs on a single NVIDIA GTX 3090 GPU and then select the model with the lowest validation EER for evaluation.

4. Results

4.1. Impacts of feature pyramid module

In this section, we study the impacts of FPM by comparing with concatenating the outputs directly under EM-Softmax and AM-Softmax. We also remove the aggregation of the outputs of each Conformer block and only use high-level feature map for final classfication to test the effect of multi-scale aggregation. As shown in Table 2, FPM has better results both in development and evaluation set no matter with AM-Softmax or EM-Softmax. FPM can improve the performance up to 16.9% in terms of min t-DCF under AM-Softmax.

Moreover, the dimension-reduced embedding visualization is shown in Figure 5. The t-distributed Stochastic Neighbour Embedding(t-SNE) is applied to evaluation dataset with and without FPM under EM-Softmax. By comparing the subfigures (a) and (b), it is shown that there are fewer spoofed speech embeddings around the bona fide speech manifold. This suggests that the embedding from our model with FPM are more discriminative.

4.2. Impacts of Elastic margin Softmax

Besides, we also study the impact of EM-Softmax by replacing it with AM-Softmax. The difference between them is that the margin penalty values are obtained from a Gaussian distribution rather than fixed and constant. Under the setting of same input features and architectures, we can see from the Table 2 that EM-Softmax surpasses AM-Softmax no matter with

Table 2: *Ablation study of MFA-Conformer on the ASVspoof 2019 logical access development and evaluation set.*

Aggregation	Dev Set		Eval Set	
	EER(%)	min t-DCF	EER(%)	min t-DCF
FPM+EM-Softmax	0.20	0.005	1.65	0.047
w/o FPM+EM-Softmax	0.58	0.019	1.84	0.054
w/o FPM+AM-Softmax	0.70	0.018	2.29	0.065
w/o concat+EM-Softmax	0.35	0.010	2.76	0.073
FPM+AM-Softmax	0.51	0.014	2.08	0.054

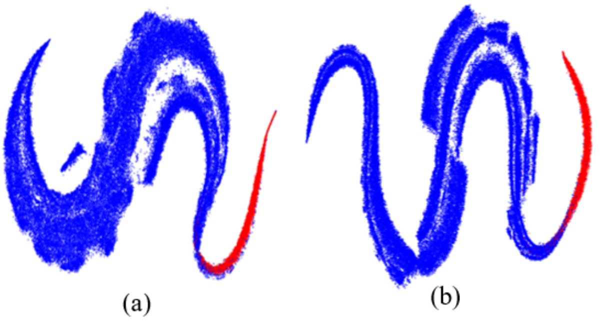

Figure 5: *Feature embedding visualization of FPM and w/o FPM. Red:bona fide speech; Blue:Spoofing attacks. (a): w/o FPM+Elastic-Softmax. (b):FPM+Elastic-Softmax.*

or without the FPM and the relative improvement on EER is up to 21.2%. We are very interested in OC-Softmax in[23] and we take it as baseline and also do a comparision with it. In Table 3, for the individual attacks, EM-Softmax has good performance in gerneral. Among all the attacks, A17 is the most difficult to detect. Compared with OC-Softmax, our system has made some improvement.

Table 3: *EER(%) and min t-DCF for individual attacks on the ASVspoof 2019 logical access evaluation and development set*

Attacks	ResNet18-OC-Softmax[23]	MFA-EM-Softmax
A07	**0.12**	0.14
A08	**0.18**	2.20
A09	0.12	**0.02**
A10	1.14	**0.91**
A11	0.12	**0.04**
A12	0.47	**0.20**
A13	0.22	**0.19**
A14	0.69	**0.47**
A15	1.40	**0.39**
A16	0.33	**0.14**
A17	9.22	**6.09**
A18	**0.90**	1.30
A19	0.90	**0.75**

4.3. Comparison of Different Models

Table 4 shows the results of the baseline system that use OC-Softmax and other systems that use various CNNs and several Conformers on the ASVSpoof2019 LA evaluation set. We also report some state-of-the-art models. In evaluation set with many unseen attacks, our proposed single system can achieve EER of 1.65%, which is a good result in single systems.

Table 4: *EER(%) and min t-DCF for different backbone networks on the ASVspoof 2019 logical access evaluation set.*

Models	Performances	
	EER(%)	min t-DCF
Res2Net[10]	2.869	0.0786
OC-Softmax[23]	2.19	0.059
Res2NEet34-Conformer[29]	1.85	0.06
ResNet-LMCL[22]	1.81	0.052
ASSERT[30]	6.70	0.155
LCNN-LSTM[31]	1.92	0.0524
LCNN[6]	1.84	0.0510
CGCNN[7]	3.56	0.1118
FG-LCNN[8]	4.07	0.102
LCNN-DA[9]	2.76	0.0777
Capsule Network[32]	1.97	0.0538
Attention[33]	1.87	0.051
Raw PC-DARTS[34]	1.77	0.0517
GAT[17]	1.68	0.0476
Ours	**1.65**	**0.047**

5. Conclusions

In this work, we propose to apply a Feature Pyramid Conformer and Elastic Margin Softmax in synthetic speech detection. Conformer can capture both local information and global information which is helpful in learning discriminative cues between spoofing and bonafide speech. By aggregating all level feature maps using feature pyramid module, the results become better. For the loss functions, we relax the fixed margin from a Gaussian distribution which improve the discriminative and generalizability of the model. Our study shows that the proposed model can enhance the robustness and generalization of the model against unknown spoofing attacks. Without any data augmentation of the ASVspoof 2019 LA scenario, the system has better performance than most existing single systems. The future work will focus on extending the studies to replay attack detection and data augmentation methods.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China [grant number 61771196].

7. References

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[2] 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.

[3] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, “Asvspoof 2017 version 2.0: metadata analysis and baseline enhancements,” in *Odyssey 2018-The Speaker and Language Recognition Workshop*, 2018.

[4] 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.

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

[6] 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.

[7] 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 *Proc.INTERSPEECH*, 2019, pp. 1038–1042.

[8] 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.

[9] X. Ma, T. Liang, S. Zhang, S. Huang, and L. He, “Improved lightcnn with attention modules for asv spoofing detection,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[10] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, “Replay and synthetic speech detection with res2net architecture,” in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2021, pp. 6354–6358.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[12] 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.

[13] 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.

[14] B. Chettri, T. Kinnunen, and E. Benetos, “Subband modeling for spoofing detection in automatic speaker verification,” *arXiv preprint arXiv:2004.01922*, 2020.

[15] 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.

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

[17] H. Tak, J.-w. Jung, J. Patino, M. Todisco, and N. Evans, “Graph attention networks for anti-spoofing,” *arXiv preprint arXiv:2104.03654*, 2021.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[19] Z. Gao, Y. Song, I. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, “Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system.” in *INTERSPEECH*, 2019, pp. 361–365.

[20] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, “Deep speaker embedding learning with multi-level pooling for text-independent speaker verification,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6116–6120.

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

[22] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, “Generalization of audio deepfake detection.” in *Odyssey*, 2020, pp. 132–137.

[23] Y. Zhang, F. Jiang, and Z. Duan, “One-class learning towards synthetic voice spoofing detection,” *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[24] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, “Elasticface: Elastic margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1578–1587.

[25] M. Kotti, E. Benetos, and C. Kotropoulos, “Computationally efficient and robust bic-based speaker segmentation,” *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 5, pp. 920–933, 2008.

[26] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, “Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification,” *arXiv preprint arXiv:2203.15249*, 2022.

[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] L. Wang, B. Yeoh, and J. W. Ng, “Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture,” *arXiv preprint arXiv:2210.03581*, 2022.

[30] 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.

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

[32] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, “A capsule network based approach for detection of audio spoofing attacks,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6359–6363.

[33] H. Ling, L. Huang, J. Huang, B. Zhang, and P. Li, “Attention-based convolutional neural network for asv spoofing detection.” in *Interspeech*, 2021, pp. 4289–4293.

[34] W. Ge, J. Patino, M. Todisco, and N. Evans, “Raw differentiable architecture search for speech deepfake and spoofing detection,” *arXiv preprint arXiv:2107.12212*, 2021.