One-Class Learning Towards Synthetic Voice Spoofing Detection

You Zhang, *Student Member, IEEE*, Fei Jiang, and Zhiyao Duan, *Member, IEEE*

***Abstract*—Human voices can be used to authenticate the identity of the speaker, but the automatic speaker verification (ASV) systems are vulnerable to voice spoofing attacks, such as impersonation, replay, text-to-speech, and voice conversion. Recently, researchers developed anti-spoofing techniques to improve the reliability of ASV systems against spoofing attacks. However, most methods encounter difficulties in detecting unknown attacks in practical use, which often have different statistical distributions from known attacks. Especially, the fast development of synthetic voice spoofing algorithms is generating increasingly powerful attacks, putting the ASV systems at risk of unseen attacks. In this work, we propose an anti-spoofing system to detect unknown synthetic voice spoofing attacks (i.e., text-to-speech or voice conversion) using one-class learning. The key idea is to compact the bona fide speech representation and inject an angular margin to separate the spoofing attacks in the embedding space. Without resorting to any data augmentation methods, our proposed system achieves an equal error rate (EER) of 2.19% on the evaluation set of ASVspoof 2019 Challenge logical access scenario, outperforming all existing single systems (*i.e.*, those without model ensemble).**

***Index Terms*—Anti-spoofing, one-class classification, feature learning, generalization ability, speaker verification.**

I. Introduction

SPEAKER verification plays an essential role in biometric authentication; it uses acoustics features to verify whether a given utterance is from the target person [1]. However, ASV systems can be fooled by spoofing attacks, such as impersonation (mimics or twins), replay (pre-recorded audio), text-to-speech (converting text to spoken words), and voice conversion (converting speech from source speaker to target speaker) [2], [3]. Among them, synthetic voice attacks (including text-to-speech (TTS) and voice conversion (VC)) are posing increasingly more threats to speaker verification systems due to the fast development of speech synthesis techniques [4], [5].

To improve the spoofing-robustness of speaker verification systems, stand-alone anti-spoofing modules are developed to detect spoofing attacks. The ASVspoof challenge series [3], [6], [7] has been providing datasets and metrics for anti-spoofing speaker verification research.

In this paper, we focus on anti-spoofing synthetic speech attacks, i.e., discriminating bona fide speech from those generated by TTS and VC algorithms. Traditional methods pay much attention to feature engineering, where good performance has been shown with hand-crafted features such as Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) [8], Linear Frequency Cepstral Coefficients (LFCC) [9], and Constant-Q Cepstral Coefficients (CQCC) [10]. As for the back-end classifier, Gaussian Mixture Model (GMM) is used in traditional methods [8]–[12]. Zhang *et al.* [13] investigated deep learning models for anti-spoofing and proved that combinations of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can improve system robustness. Monteiro *et al.* [14] proposed to adopt the deep residual network (ResNet) with temporal pooling. Chen *et al.* [15] proposed a system which employs the ResNet with large margin cosine loss and applied frequency mask augmentation. Gomez-Alanis *et al.* [16] adapted a light convolutional gated RNN architecture to improve the long-term dependency for spoofing attacks detection. Wu *et al.* [17] proposed a feature genuinization based light CNN system that outperforms other single systems for detection of synthetic attacks. Aravind *et al.* [18] explored transfer learning approach with a ResNet architecture. To further improve the performance, researchers introduced model fusion based on sub-band modeling [19] and different features [20]–[22] at the cost of increased model complexity.

While much progress has been made, existing methods generally suffer from generalization to unseen spoofing attacks in the test stage [2], [4]. We argue that this is because most methods formulate the problem as binary classification of bona fide and synthetic speech, which intrinsically assumes the same or similar distributions between training and test data for both classes. While this assumption is reasonable for the bona fide class given a big training set with diverse speakers, it is hardly true for the fake class. Due to the development of speech synthesis techniques, the synthetic spoofing attacks in the training set may never be able to catch up with the expansion of the distribution of spoofing attacks in practice.

This distribution mismatch between training and test for the fake class actually makes the problem a good fit for one-class classification. In the one-class classification setup [23], there is a target class that does not have this distribution mismatch problem, while for the non-target class(es), samples in the training set are either absent or statistically unrepresentative. The key idea of one-class classification methods is to capture the target class

Manuscript received February 8, 2021; revised April 9, 2021; accepted April 25, 2021. Date of publication April 28, 2021; date of current version May 20, 2021. This work was supported by the National Science Foundation under Grant 1741472 and by the Voice Biometrics Group. The associate editor coordinating the review of this manuscript and approving it for publication Dr. Xiangui Kang. *(Corresponding author: Zhiyao Duan.)*

You Zhang and Zhiyao Duan are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA (e-mail: you.zhang@rochester.edu; zhiyao.duan@rochester.edu).

Fei Jiang is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA, and also with the Beijing Institute of Technology, Beijing 100081, China (e-mail: flyjiang92@gmail.com).

Digital Object Identifier 10.1109/LSP.2021.3076358

distribution and set a tight classification boundary around it, so that all non-target data would be placed outside the boundary.

Recently, the one-class learning idea has been successfully introduced into image forgery detection [24]–[26]. For voice spoofing detection, Alegre *et al.* [27] employed a one-class support vector machine (OC-SVM) trained only on bona fide speech to classify local binary patterns of speech cepstrograms, showing the potential of one-class classification approach. It did not make use of any information from spoofing attacks. Villalba *et al.* [28] proposed to fit OC-SVM with DNN extracted speech embeddings of the bona fide class for the ASVspoof 2015 Challenge. Although the method uses OC-SVM to learn a compact boundary for the bona fide class, the embedding space is still learned through binary classification. In other words, the embedding space for drawing the classification boundary may not benefit one-class classification.

In this paper, we formulate the speaker verification anti-spoofing problem as one-class feature learning to improve the generalization ability. The target class refers to bona fide speech and the non-target class refers to spoofing attacks. We propose a loss function called one-class softmax (OC-Softmax) to learn a feature space in which the bona fide speech embeddings have a compact boundary while spoofing data are kept away from the bona fide data by a certain margin. Our proposed method, without resorting to any data augmentation, outperforms all existing single systems (those without ensemble learning) on the ASVspoof 2019 LA dataset, and ranks between the second and third places among all participating systems.

II. Method

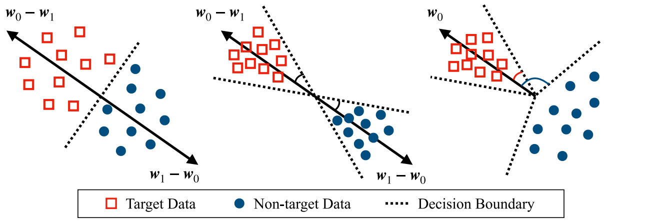
Typically, for deep learning-based voice spoofing detection models, the speech features are fed into a neural network to calculate an embedding vector for the input utterance. The objective of training this model is to learn an embedding space in which the bona fide voices and spoofing voices can be well discriminated. The embedding would be further used for scoring the confidence of whether the utterance belongs to bona fide speech or not. To the best of our knowledge, all the previous voice spoofing detection systems learn speech embedding using a binary classification loss function. As discussed in Section I, this may limit the generalization ability to unknown attacks as spoofing algorithms evolve. In this section, we first briefly introduce and analyze the widely used binary classification loss functions, then propose our one-class learning loss function for voice spoofing detection.

A. Preliminary: Binary Classification Loss Functions

The original Softmax loss for binary classification can be formulated as

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^T\boldsymbol{x}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{(\boldsymbol{w}_{1-y_i}-\boldsymbol{w}_{y_i})^T\boldsymbol{x}_i}\right), \quad (1)$$

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 respectively, $\boldsymbol{w}_0, \boldsymbol{w}_1 \in \mathbb{R}^D$ are the weight vectors of the two classes, and N is the number of samples in a mini-batch.



(a) Original Softmax (b) AM-Softmax (c) OC-Softmax (Proposed)

Fig. 1. Illustration of the original Softmax and AM-Softmax for binary classification, and our proposed OC-Softmax for one-class learning. (The embeddings and the weight vectors shown are non-normalized).

AM-Softmax [29] improves upon this by introducing an angular margin to make the embedding distributions of both classes more compact, around the weight difference vector's two directions:

$$\mathcal{L}_{AMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}}{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}+e^{\alpha\hat{\boldsymbol{w}}_{1-y_i}^T\hat{\boldsymbol{x}}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha\left(m-(\hat{\boldsymbol{w}}_{y_i}-\hat{\boldsymbol{w}}_{1-y_i})^T\hat{\boldsymbol{x}}_i\right)}\right), \quad (2)$$

where α is a scale factor, m is the margin for cosine similarity, and $\hat{\boldsymbol{w}}$, $\hat{\boldsymbol{x}}$ are normalized $\boldsymbol{w}$ and $\boldsymbol{x}$ respectively.

B. Proposed Loss Function for One-Class Learning

According to the formulae of Softmax and AM-Softmax in (1) and (2), for both loss functions, the embedding vectors of the target and non-target class tend to converge around two opposite directions, i.e., $\boldsymbol{w}_0 - \boldsymbol{w}_1$ and $\boldsymbol{w}_1 - \boldsymbol{w}_0$, respectively. This is shown in Fig. 1(a), (b). For AM-Softmax, the embeddings of both target and non-target class are imposed with an identical compactness margin m. The larger m is, the more compact the embeddings will be.

In voice spoofing detection, it is reasonable to train a compact embedding space for bona fide speech. However, if we also train a compact embedding space for the spoofing attacks, it may overfit known attacks. To address this issue, we propose to introduce two different margins for better compacting the bona fide speech and isolating the spoofing attacks. The proposed loss function One-class Softmax (OC-Softmax) is denoted as

$$\mathcal{L}_{OCS} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha(m_{y_i}-\hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i)(-1)^{y_i}}\right). \quad (3)$$

Note that only one weight vector $\boldsymbol{w}_0$ is used in this loss function. The $\boldsymbol{w}_0$ refers to the optimization direction of the target class embeddings. The $\boldsymbol{w}_0$ and $\boldsymbol{x}$ are also normalized as in AM-Softmax. Two margins $(m_0, m_1 \in [-1,1], m_0 > m_1)$ are introduced here for bona fide speech and spoofing attacks respectively, to bound the angle between $\boldsymbol{w}_0$ and $\boldsymbol{x}_i$, which is denoted by θ_i. When $y_i = 0$, m_0 is used to force θ_i to be smaller than $\arccos m_0$, whereas when $y_i = 1$, m_1 is used to force θ_i to be larger than $\arccos m_1$. As shown in Fig. 1 (c), a small $\arccos m_0$ can make the target class concentrate around

TABLE I
SUMMARY OF THE ASVSPOOF 2019 LA DATASET

	Bona fide	Spoofed	
	# utterance	# utterance	attacks
Training	2,580	22,800	A01 - A06
Development	2,548	22,296	A01 - A06
Evaluation	7,355	63,882	A07 - A19

TABLE II
RESULTS ON THE DEVELOPMENT AND EVALUATION SETS OF THE ASVSPOOF 2019 LA SCENARIO USING LOSS FUNCTIONS IN SECTION II

Loss	**Dev Set**		**Eval Set**	
	EER (%)	min t-DCF	EER (%)	min t-DCF
Softmax	0.35	0.010	4.69	0.125
AM-Softmax	0.43	0.013	3.26	0.082
OC-Softmax	0.20	0.006	**2.19**	**0.059**

TABLE III
EER (%) PERFORMANCE COMPARISON OF LOSS FUNCTIONS ON INDIVIDUAL ATTACKS OF THE EVALUATION SET (ALL UNSEEN FROM DEVELOPMENT) OF THE ASVSPOOF 2019 LA SCENARIO. (* MEANS THE INDIVIDUAL EER IS STATISTICALLY SIGNIFICANTLY DIFFERENT FROM OC-SOFTMAX, WITH A 99% CONFIDENCE INTERVAL)

Attacks	Softmax	AM-Softmax	OC-Softmax
A07	0.37*	0.22	0.12
A08	0.01	0.06	0.18
A09	0.02*	0.02*	0.12
A10	1.18	0.63*	1.14
A11	0.37	0.02*	0.12
A12	0.43	0.53	0.47
A13	0.69*	0.27	0.22
A14	4.98*	0.51	0.69
A15	5.43*	0.69*	1.40
A16	0.22	0.51	0.33
A17	23.48*	13.45*	9.22
A18	0.20*	4.27*	0.90
A19	1.34*	0.86	0.90

the weight vector $\boldsymbol{w}_0$, whereas a relatively large $\arccos m_1$ can push the non-target data to be apart from $\boldsymbol{w}_0$.

III. EXPERIMENTS

A. Dataset

The ASVspoof 2019 challenge provides a standard database [30] for anti-spoofing. The LA subset of the provided dataset includes bona fide 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. In the evaluation set, there are 11 unknown attacks (A07-A15, A17, A18) including combinations of different TTS and VC attacks. The evaluation set also includes two attacks (A16, A19) which use the same algorithms as two of the attacks (A04, A06) in the training set but trained with different data. Details can be found in Table I.

B. Evaluation Metrics

To evaluate the performance of the anti-spoofing system, we take note of the output score of the anti-spoofing system. The output of the anti-spoofing system is called a countermeasure (CM) score, and it indicates the similarity of the given utterance with bona fide speech. For systems trained with Softmax or AM-Softmax, the output CM score is the cosine similarity between the speech embedding $\boldsymbol{x}_i$ and the weight vector $\boldsymbol{w}_0 - \boldsymbol{w}_1$. For our proposed system with OC-Softmax, the CM score is the cosine similarity of $\boldsymbol{x}_i$ and $\boldsymbol{w}_0$. Equal Error Rate (EER) is calculated by setting a threshold on the CM decision score such that the false alarm rate is equal to the miss rate. The lower the EER is, the better the anti-spoofing system is at detecting spoofing attacks. The tandem detection cost function (t-DCF) [31] is a new evaluation metric adopted in the ASVspoof 2019 challenge. While EER only evaluates the performance of the anti-spoofing system, the t-DCF assesses the influence of anti-spoofing systems on the reliability of an ASV system. The ASV system is fixed to compare different anti-spoofing systems. The lower the t-DCF is, the better reliability of ASV is achieved.

C. Training Details

We extract 60-dimensional LFCCs from the utterances with the MATLAB implementation provided by the ASVspoof 2019 Challenge organizers.[1] The frame size is 20 ms and the hop size is 10 ms. To form batches, we set 750 frames as the fixed length and use repeat padding for short trials, and we randomly choose a consecutive piece of frames and discard the rest for long trials.

We adopt the network architecture adapted from [14]. The architecture is based on deep residual network ResNet-18 [32], where the global average pooling layer is replaced by attentive temporal pooling. The architecture takes the extracted LFCC features as input and outputs the confidence score to indicate the classification result. We use the intermediate output before the last fully connected layer as the embedding for the speech utterances, where 256 is the embedding dimension. For the hyper-parameters in the loss functions, we set $\alpha = 20$ and $m = 0.9$ for AM-Softmax; we set $\alpha = 20$, $m_0 = 0.9$ and $m_1 = 0.2$ for the proposed OC-Softmax.

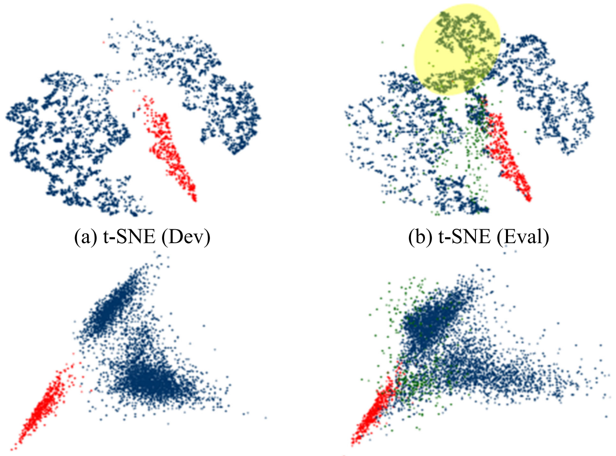
We implement our model with PyTorch.[2] We use Adam optimizer with the β_1 parameter set to 0.9 and the β_2 parameter set to 0.999 to update the weights in the ResNet model. We use Stochastic Gradient Descent (SGD) optimizer for the parameters in the loss functions. The batch size is set to 64. The learning rate is initially set to 0.0003 with 50% decay for every 10 epochs. We trained the network for 100 epochs on a single NVIDIA GTX 1080 Ti GPU. Then we select the model with the lowest validation EER for evaluation.

D. Results

1) Evaluation of Proposed Loss Function: To demonstrate the effectiveness of the one-class learning method, we compared our proposed OC-Softmax with the conventional binary classification loss functions, under the setting of the same input features and models. The performance of the system trained with different loss functions is compared in Table II on both the development and evaluation sets of ASVspoof 2019 LA scenario. We also compare the performance on individual unseen attacks in the evaluation set in Table III.

[1] [Online]. Available: https://www.asvspoof.org

[2] [Online]. Available: https://github.com/yzyouzhang/AIR-ASVspoof



(a) t-SNE (Dev)

(b) t-SNE (Eval)

(c) PCA (Dev)

(d) PCA (Eval)

Fig. 2. Feature embedding visualization of our proposed loss. Red: Bona fide Speech; Blue: Spoofing Attacks except A17. Green: A17 attack. Highlighted part are the unknown attacks with different distributions.

TABLE IV
PERFORMANCE COMPARISON WITH EXISTING SINGLE SYSTEMS ON THE EVALUATION SET OF THE ASVSPOOF 2019 LA SCENARIO

System	EER (%)	min t-DCF
CQCC + GMM [3]	9.57	0.237
LFCC + GMM [3]	8.09	0.212
Chettri et al. [22]	7.66	0.179
Monterio et al. [14]	6.38	0.142
Gomez-Alanis et al. [16]	6.28	-
Aravind et al. [18]	5.32	0.151
Lavrentyeva et al. [21]	4.53	0.103
ResNet + OC-SVM	4.44	0.115
Wu et al. [17]	4.07	0.102
Tak et al. [19]	3.50	0.090
Chen et al. [15]	3.49	0.092
Proposed	**2.19**	**0.059**

The three losses perform similarly on the development set, showing that they have good discrimination ability to detect known attacks. For the evaluation set, where all the attacks are not included in the development set, our one-class learning with the proposed OC-Softmax surpasses the binary classification losses (Softmax and AM-Softmax). The relative improvement on EER is up to 33%. As for individual attacks, our system with OC-Softmax achieves universally good performance over all but A17 attacks, showing the efficacy of one-class learning to detect unknown attacks. A17 is the most difficult attack in the evaluation set [3], but our proposed system shows a significant improvement. We conduct a statistical significance test based on [33], and in Table III we mark the individual EERs that show significant difference on Softmax and AM-Softmax, comparing with OC-Softmax.

Moreover, the dimension-reduced embedding visualization is shown in Figure 2. The same t-distributed Stochastic Neighbor Embedding (t-SNE) and Principle Component Analysis (PCA) projections are applied to development and evaluation datasets of ASVspoof 2019 LA scenario. In other words, the visualizations of the two sets use the same coordinating systems. The t-SNE subfigures show that the bona fide speech has the same distribution in both sets while unknown attacks in the evaluation set show different distributions from the known attacks in the development set. This suggests that the bona fide class is well characterized by the instances in the training data, but the spoofing attacks in the training data cannot form a statistical representation of the unknown attacks, hereby verifies our problem formulation in Section I. Nevertheless, the unknown attacks that appear in the top cluster (highlighted) and especially the A17 attack (green) on the top right subfigure are successfully separated from the bona fide speech cluster, showing a good generalization ability of our system.

We further verify the one-class idea with PCA visualization of the learned embedding. In the bottom left subfigure, the embeddings of the bona fide speech are compact, and an angular margin is injected between bona fide and spoofing attacks, thanks to the linearity of PCA, verifying our assumption in Figure 1 (c). The bottom right figure shows that when encountering unknown spoofing attacks, the angle is still maintained, and the embeddings for the unknown attacks are still mapped to the angularly separate space. This shows the effectiveness of our proposed OC-Softmax loss.

2) Comparison With Other Systems: To demonstrate the superiority of our proposed method, we compared our system with other existing single systems (no model fusion) without data augmentation in Table IV and also with the leaderboard of the ASVspoof 2019 Challenge for LA scenario [3].

For all methods in Table IV with a reference, we obtained their results from their papers. Some of them participated in the ASVspoof 2019 LA challenge and reported better results with model ensembles and data augmentation, but we only compare with their single system version without data augmentation. It is noted that "ResNet + OC-SVM" was adapted from [28], which is the only existing one-class classification method; We replaced their DNN with our ResNet and run the experiments on the ASVspoof 2019 LA dataset for a fair comparison. It can be seen that our proposed system significantly outperforms all of the other single systems.

In fact, on the leader board of the ASVspoof 2019 Challenge for LA scenario [3], our system would rank between the second (EER 1.86%) [21] and the third (EER 2.64%) [22] among all the systems, even though the top three methods all used model fusion.

IV. CONCLUSION

In this work, we proposed a voice spoofing detection system based on one-class learning to enhance the robustness of the model against unknown spoofing attacks. The proposed system aims to learn a speech embedding space in which bona fide speech has a compact distribution while spoofing attacks reside outside by an angular margin. Experiments showed that the proposed loss outperforms the original Softmax and AM-Softmax that formulate anti-spoofing as a conventional binary classification problem. The proposed system also outperforms all existing single systems (no model fusion) without data augmentation of the ASVspoof 2019 Challenge LA scenario, and ranks between the second and the third among all participating systems. For future work, we would like to extend our method to detecting other multimedia forgeries.

REFERENCES

[1] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proc. Elmar-2004. 46th Int. Symp. Electron. Mar.*, 2004, pp. 184–193.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[3] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[4] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of ASVspoof challenges," *APSIPA Trans. Signal Inf. Process.*, vol. 9, p. e2, 2020.

[5] R. K. Das *et al.*, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 99–120.

[6] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2037–2041.

[7] T. Kinnunen *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017, pp. 2–6. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1111

[8] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2062–2066.

[9] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2087–2091.

[10] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey*, vol. 45, 2016, pp. 283–290.

[11] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2092–2096.

[12] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Inf. Foren. Security*, vol. 10, no. 4, pp. 810–820, Apr. 2015.

[13] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 684–694, Jun. 2017.

[14] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, 2020, Art. no. 101096.

[15] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 132–137.

[16] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Interspeech*, 2019, pp. 1068–1072.

[17] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Proc. Interspeech*, 2020, pp. 1101–1105.

[18] P. Aravind *et al.*, "Audio spoofing verification using deep convolutional neural networks by transfer learning," 2020, *arXiv:2008.03464*.

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

[20] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 2119–2123.

[21] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. Interspeech*, 2019, pp. 1033–1037.

[22] 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," in *Proc. Interspeech*, 2019, pp. 1018–1022. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2505

[23] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Proc. Ir. Conf. Artif. Intell. Cogn. Sci*. Berlin, Germany: Springer, 2009, pp. 188–197.

[24] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 656–657.

[25] Y. Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2020, pp. 1–9.

[26] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 667–684.

[27] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. IEEE 6th Int. Conf. Biometrics: Theory, Appl. Syst.*, 2013, pp. 1–8.

[28] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2067–2071.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[30] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthetized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, pp. 101–114, Nov. 2020.

[31] T. Kinnunen *et al.*, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 312–319. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-44

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[33] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, 2004, pp. 237–244.