Robust Deepfake Audio Detection via Bi-level Optimization

1st Menglu Li
Faculty of Electrical and Computer Engineering
Toronto Metropolitan University, Canada
menglu.li@torontomu.ca

2nd Yasaman Ahmadiadli
Faculty of Electrical and Computer Engineering
Toronto Metropolitan University, Canada
yahmadiadli@torontomu.ca

3rd Xiao-Ping Zhang
Faculty of Electrical and Computer Engineering
Toronto Metropolitan University, Canada
xzhang@ee.torontomu.ca

2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP) | 979-8-3503-3893-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/MMSP59012.2023.10337724

***Abstract*—ASVspoof Challenges have been launched to motivate research on Deepfake audio detection due to its threats to society. However, the state-of-the-art detection models produce an unsatisfactory performance on the Speech Deepfake (DF) of the challenge. The DF subset includes spoofed audio from various sources, which can better reflect the robustness of the detector. In this paper, we propose a novel detection architecture to improve the robustness and generalization ability in two ways. The first way is aggregating both learned embeddings and hand-crafted features to obtain more generalizable representations for Deepfake audio. Our second contribution is formulating the training process a bi-level optimization problem to make use of the knowledge of different Deepfake generation methods. Evaluations of our proposed method provide the best detection output reported in the literature as a single system without the help of ensemble modeling and data augmentation.**

***Index Terms*—Deepfake, Audio Deepfake Detection, Anti-Spoofing, ASVspoof2021**

I. Introduction

Online communication is prone to misinformation, hoax, and digital forgery. The use of Deep Learning techniques for spoofed audio production has resulted in Deepfake audio. With the advancements to generate spoofed audio, Deepfakes are spreading rapidly across the world. The dangerous impact of Deepfake audio throughout societies and countries calls for efficient spoof detection approaches in Automatic Speaker Verification (ASV) tasks.

The ASVspoof Challenges are hosted regularly to inspire academicians to propose cutting-edge research spoofed audio detection [1]. In the ASVspoof2019 challenge [2], two main categories were included namely, Logical Access (LA) and Physical Access (PA). The LA category mainly focuses on Text-to-Speech (TTS) and Voice Conversion (VC) generation techniques while, the PA category comprises replay attacks. In the ASVspoof2021 Challenge [3], the Speech Deepfake (DF) category has been added as a new task, which aims to motivate the robustness of spoofing detection solutions. The full DF evaluation database contains spoofed audio from various sources in different storage formats, which are generated with more than 100 spoofing algorithms that differ from those seen in the training set [4]. However, the state-of-the-art detectors that achieve top performance in the LA track result in a high degree of overfitting in the DF track [3]. The reason behind this is that most existing detection models are built based on the training set that contains existing spoofing methods without utilizing the knowledge from diverse unknown generating methods. Furthermore, all the top-5 performing submissions of the ASVspoof2021 DF track use score averaging to build a fusion system and apply data augmentation (DA) techniques instead of single systems [4].

To address this issue, our work focuses on feature extraction and optimization techniques to make the single model detector more generalized and robust. Our model extracts both learned features and hand-crafted features from raw audio inputs. We also formulate the training process into a bi-level optimization problem, which optimizes the model parameters to make use of the knowledge to the utmost from various spoofing generation algorithms. The main contributions of this paper include:

- We propose a novel Deepfake audio detection model with bi-level optimization to increase the robustness and generalization ability of the detector. It provides the best detection result on the ASVspoof2021 DF evaluation dataset as a single model without data augmentation.
- We propose a Transformer-based classifier trained on the double components of feature extraction. Combining hand-crafted features and learned features produces a

more generalizable representation for Deepfake audios with varying characteristics.
- We demonstrate the effectiveness of bi-level optimization in detecting fake audio generated by unseen algorithms with a systematic experiment.

II. Related work

The work in [5] demonstrated the effectiveness of the SincNet filters as a novel CNN-based layer based on parametrized sinc functions to achieve faster convergence and better performance in speaker and speech recognition. Tak et al. [6] first adopted SincNet to process the raw audio data and produce deep embeddings in an End-to-End Deepfake audio detector. Then, [7] continues integrating the SincNet component with graph attention networks (GAT). Both models become the most well-known reproducible detector for Deepfake audio. Besides this learned feature, the hand-crafted acoustic and perceptual features, such as Mel-frequency Cepstral Coefficients (MFCC) [8], [9], Constant Q Cepstral Coefficients (CQCC) [10], [11], and Perceptual Linear Predictive Coefficient (PLP) [12], have been widely used in the Deepfake detectors to extract audio representations. Our assumption is that using either hand-crafted features or learned embeddings only, will form the Deepfake audio detection model based on the training set highly where the detection model can not achieve a more generalized performance in detecting fake audios generated by unseen methods. Therefore, we propose to integrate both forms of feature extraction with bi-level optimization technique to further enhance a robust detection performance as a single model.

Bi-level optimization is a nested optimization problem where a set of variables involved in the outer objective function are obtained by solving another inner objective optimization problem [13]. The inner problem works as a constraint to the outer objective problem, such that only the optimal solution to the inner problem can be a feasible optimal solution to the outer problem [14]. It has shown its effectiveness in the area of data augmentation learning [14], molecular conformation generation [15], and image inpainting detection [16]. It has never been applied in the audio domain, and we believe that bi-level optimization has the potential to obtain more generalizable parameters to detect unseen attacks.

In recent years, LCNN [17], [18] and ResNet [19], [20] are the most popular choices for the classifier producing significant performance in the LA track of ASVspoof challenges. Motivated by the success of the Transformers in the natural language processing [21] and vision domain [22], the Transformer encoder has been adopted as the classifier in detecting synthesized speeches [23], [24]. However, the existing Transformer-based models do not show sufficient generalization ability to identify spoofing audios from different sources or against various codecs. We want to investigate whether the bi-level optimization technique can be integrated into the Transformer-based model to gain robustness. The new contribution of integrating both feature components as inputs to the Transformer encoder is described in the next section.

III. Proposed Model

The overall workflow of the proposed detector is illustrated in Figure 1, which consists of three main components. The Transformer-based detection classifier is trained on both learned and hand-crafted features. A detailed explanation of each component is provided in the following subsections. We propose to formulate the training process of the proposed detector as a bi-level optimization problem, as Figure 2 shows, for robustness improvement.

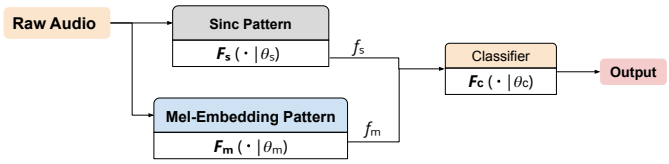

Fig. 1. The illustration of components for the proposed model

A. Sinc Pattern

The role of this component is to extract a set of learned embedding by applying SincNet filters. We implement a Sinc convolution layer with 128 SincNet filters. For each filter, the cut-in and cut-off frequencies are learnt from the training data. Our technique to interpret the output of this Sinc layer is to consider it as a two-dimensional representation with one channel. Therefore, instead of applying a set of Conv layers or ResNet blocks, we split this 2-D output into a sequence of equal-sized patches and flattened each patch into a 1-D embedding, whose aim is to retain spatial information. Then, the linear projection layer is applied to reduce the embedding length for the transformer encoder. The resulting high-level embeddings, f_s, are the final feature representation for the Sinc Pattern component.

B. Mel-Embedding Pattern

In addition to the SincNet-based feature, we also obtain a separate set of hand-crafted features. To prevent overfitting on known attacks, the mel-spectrogram, as a hand-crafted feature, is added to provide additional acoustic information on unknown attacks. This component takes raw audio as input and extracts the corresponding Mel-spectrogram, which describes the frequency and amplitude of the raw audio in a Mel-scale and becomes more consistent with the human hearing system. To match the shape of the feature representation from the previous Sinc Pattern component, the Mel-spectrograms are passed into three convolutional layers with batch-Norm and ReLU functions to obtain an embedding representation named f_m.

C. Classifier Component

From the previous two components, we project both learned features and hand-crafted features into the same dimension and then concatenate these two sets of features and pass them to the Transformer-based classifier. Positional encoding is applied to this concatenated feature embedding, in order to aggregate

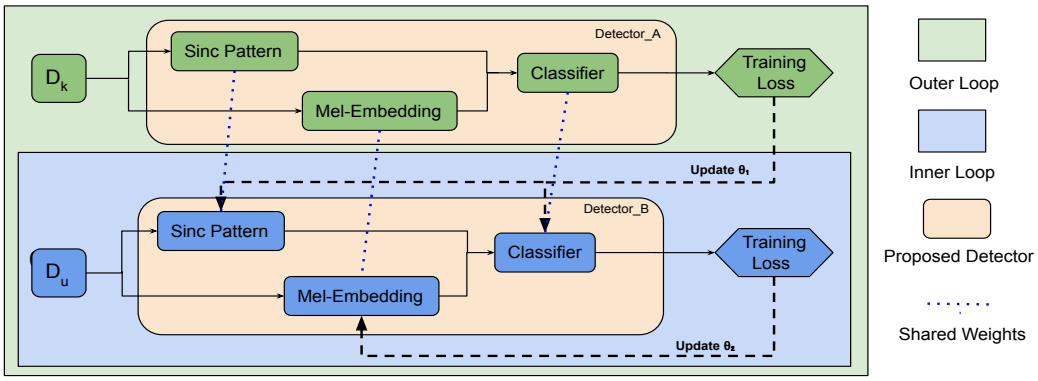

Fig. 2. **Model training via bi-level optimization.** The training dataset is split into two subsets, D_k and D_u, and passed into two identical detectors. In each epoch, the parameters of the Mel-Embedding component, θ_2, are trained on D_u in the inner loop. In the outer loop, the rest of the parameters in the detector are trained on D_k based on optimal θ_2^*

the information of input order. Since audio Deepfake detection is a binary classification task, we only utilize the encoder component of the Transformer. The output of the Transformer encoder is mapped into a binary label using a linear layer as the detection result.

D. Training Process with bi-level Optimization

The training process for our detection model can be formalized as:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim D}\, \mathcal{L}(y_{true}, y_{pred}|\theta), \tag{1}$$

where x is input audio, y is its corresponding label, D stands for the training dataset, θ denotes the parameters of the entire model, and $\mathcal{L}$ is the loss function that compares the difference between the predicted label, y_{pred}, and true label, y_{true}. The standard training process, which adopts the whole training dataset to update the parameters of the entire model together as (1) indicates, may only utilize the knowledge of Deepfake audio in the training set, without being generalizable to unseen spoofing attacks. To address this issue, we formulate the training process into a bi-level optimization problem to improve the robustness of the detection model to detect various attacks.

As the first step, we consider that the trainable parameters of the entire detection model, θ, consists of three sets based on different components, as Figure 1 shows. The operation of Sinc Pattern, F_s, is parameterized by θ_s; the parameters in the Mel-Embedding Pattern, F_m, are denoted by θ_m, and the transformer-based classifier,F_c, contains trainable parameters θ_c. Therefore, we obtain the following equations:

$$f_s = \boldsymbol{F_s}(x\,|\,\theta_s), \tag{2}$$

$$f_m = \boldsymbol{F_m}(x\,|\,\theta_m), \tag{3}$$

$$y_{pred} = \boldsymbol{F_c}[concat(f_s, f_m)\,|\,\theta_c], \tag{4}$$

Then, we split the entire training parameters, $\{\theta_s, \theta_m, \theta_c\}$, into two subsets as follows,

$$\theta_1 = \{\theta_s, \theta_c\},$$
$$\theta_2 = \{\theta_m\}. \tag{5}$$

We group the parameters of the SincNet features, θ_s, and the Transformer, θ_c, as θ_1, because combining the learned embedding and the attention-based classifier can extract the underlying characteristics of known spoofing attacks to the best extent. To prevent overfitting on known attacks, the Mel-spectrogram, as a hand-crafted feature, is added to provide additional acoustic information on unknown attacks, where this step forms a separate partition named θ_2.

We also introduce a new way to utilize the training dataset, D, by splitting it into two sets, D_u and D_k. We assume that the set D_k represents the audio generated by some known and typical generating methods, while the audio samples in the set D_u are generated by the diverse unknown generating algorithms that are different from D_k. We partition the training set based on the generation methods so the model can learn the discriminative information among various generation algorithms. We want to use D_k and D_u to train θ_1 and θ_2 separately.

To set up a bi-level optimization problem, we define the optimization of θ_1 to minimize the training loss on the data D_k to be the outer objective problem. The inner objective problem is to optimize θ_2 on the training dataset D_u. The formulation can be written as follows

$$\theta_1^* = \arg\min_{\theta_1} \mathbb{E}_{(x,y)\sim D_k}\, \mathcal{L}(y_{true}, y_{pred}|\theta_1, \theta_2^*) \tag{6}$$
$$s.t. \quad \theta_2^* = \arg\min_{\theta_2^*} \mathbb{E}_{(x,y)\sim D_u}\, \mathcal{L}(y_{true}, y_{pred}|\theta_1, \theta_2)$$

After initializing θ_1 and θ_2, we optimize the value of θ_2 using a standard supervised training with gradient descent on dataset D_u, whose aim is to learn the patterns of fake audio generated by the diverse methods. In order to make the optimal θ_2 work as a constraint to the outer objective problem, we then fix the θ_2 at its optimal value and update the θ_1 on the D_k set. In this case, θ_1 is optimized by aggregating the features of fake audio from both typical and diverse generation algorithms. There is still one challenge in training this derived bi-level optimization equation: finding the global optimal θ_2 throughout the entire training iterations to update

θ_1 is infeasible due to the computation cost. We address this issue by utilizing a local optimal θ_2 as the constraint to the optimization of θ_1. We set the local optimal θ_2 as obtained after each training epoch. The details are described in Algorithm 1.

Algorithm 1 The bi-level Optimization Algorithm

Require: Dataset D and two identical detection models, $Detector\_A$ and $Detector\_B$, that have the same initial values of the parameters. $Detector\_A$ has parameters $\{\theta_s^A, \theta_m^A, \theta_c^A\}$. and $Detector\_B$ is parameterized by $\{\theta_s^B, \theta_m^B, \theta_c^B\}$. Each model has a learning rate, μ_A and μ_B, respectively.

for number of epochs **do**
  Split D into D_k and D_u
  for k-th minibatch **do**
    // Update θ_s^A and θ_c^A with θ_m^A unchanged in Detector_A
    θ_m^A.requires_grad = $False$
    $\mathcal{L}_A = Detector\_A(D_k)$
    $\theta_s^A \leftarrow \theta_s^A - \mu_A \frac{\partial \mathcal{L}_A}{\partial \theta_s^A}$
    $\theta_c^A \leftarrow \theta_c^A - \mu_A \frac{\partial \mathcal{L}_A}{\partial \theta_c^A}$

    // Share the values of θ_1 in Detector_A with Detector_B
    $\theta_s^B \leftarrow \theta_s^A$
    $\theta_c^B \leftarrow \theta_c^A$

    // Update θ_m^B with θ_s^B and θ_c^B unchanged in Detector_B
    θ_s^B.requires_grad = $False$
    θ_c^B.requires_grad = $False$
    $\mathcal{L}_B = Detector\_B(D_u)$
    $\theta_m^B \leftarrow \theta_m^B - \mu_B \frac{\partial \mathcal{L}_B}{\partial \theta_m^B}$
  end for
  // Update the local optimal θ_2 to Detector_A
  $\theta_m^A \leftarrow \theta_m^B$
end for
return $\theta_s^A, \theta_m^A, \theta_c^A$

We believe that our bi-level optimization mechanism is able to make full use of the knowledge of different generation methods of Deepfake audio so that we can obtain a detector with a more generalized detection capacity.

IV. Experiment

Here we explain the dataset details along with the training configurations we use for our proposed model. We also demonstrate our detection results against other state-of-the-art on the same evaluation data.

A. Datasets and Evaluation Metrics

In order to investigate the robustness of the proposed detector, we use the audio data under Deepfake (DF) track in ASVspoof2021 dataset for the experiment. The ASVspoof2021 evaluation set is the largest and most diverse online-available dataset and it contains audio clips from other domains such as VCC2018 and VCC2020 datasets [25]. Thus, we believe the performance reported on this evaluation dataset can be the strongest evidence to demonstrate the robustness of the detectors in the literature. The training data for the DF track are the same as the ASVspoof2019 logical access (LA) set, in which the spoofed audio is generated by four TTS and two VC methods. The ratio of real speech and spoofed speech in the training set is 1:9. During training, we fix all audio to the same length of 4 seconds either by truncating the longer audio clips or concatenating the shorter audio clips repeatedly.

We adopt Equal Error Rate (EER) as the indicator to measure the detection performance. The EER corresponds to a threshold where the false positive rate and false negative rate are equal. A detection result with a lower EER score is regarded to be more accurate. The EER is also used as the official evaluation metric for the DF track of ASVspoof2021 Challenge.

B. Model and Training Configurations

In the Sinc Pattern, we use 128 band-pass filters with a length of 80 to operate on the raw waveform directly. After max pooling, we split the resulting embedding into 24x24 patches and flattened each patch to a 1D embedding of size 256. Three linear projection layers are applied to shorten the embedding length resulting in a 128x256 matrix as output for Sinc Pattern. For the Mel-Embedding Pattern, the feature is computed using the default setting in TorchAudio Python Framework with a mel-filter size of 128 and a fast Fourier Transform size of 400. Three convolutional layers project the extracted Mel-spectrogram to a deep embedding with a dimension size of 256, which is the same as Sinc embedding patches. Then, two deep representations of the audio input with the same dimension size are concatenated and passed into the Transformer encoder, which has an embedding dimension of 1024, 6 blocks with 4 heads.

We utilize the whole training set to train our model. In order to split up the training set into D_u and D_k, we simulate the technique of 3-fold cross-validation. We have six types of spoofed audio which can be divided into six groups, and then we equally spread out the bonafide audio into these six groups. Each fold contains two groups of training data. During each epoch, two folds of data are randomly selected to be the set D_k, and the rest of the training data will be the D_u set. The model was trained with Adam optimizer using a learning rate of 0.001 as μ_A, 0.005 as μ_B, and 150 epochs with a mini-batch size of 32. The model with the minimum validation loss on the development dataset is selected as the best model for evaluation during training epochs.

C. Results

The performance results are illustrated in Table I. The models B01 to B04 are the baselines of the ASVspoof2021 Challenge, and the corresponding evaluation results on the DF set are reported in [3]. All other systems in Table I are selected because they have reported the state-of-the-art performance on the ASVspoof2019 LA set, and they also have provided the source code or pre-trained models online to allow reproduction. We then either train the online source codes with the provided hyperparameters or directly apply the pre-trained models and test them on the DF track evaluation set. We observe that our proposed model outperforms the official baselines with at least a 9.56% improvement in the EER metric (20.24% cf. 22.38%). As results indicate, our

approach reports the best performance for the DF evaluation set as a single model system without data augmentation. This result emphasizes the robustness and accuracy of our proposed model on Deepfake detection since the spoofed audio samples in the DF evaluation set are generated by various algorithms that differ from the training set. Notably, the Melspectrogram-LCNN model achieves 16.96% EER with DA techniques. However, without DA, the performance of this model decreases to an EER of 21.90%, which is worse than our model by 7.6% (21.90% cf. 20.24%). We believe it is meaningful to compare models' performance without DA, which reflects the detection ability of the model architecture under the same condition of training data.

TABLE I
PERFORMANCE ON THE ASVSPOOF2021 DF EVALUATION SET FOR OUR PROPOSED MODEL AND DIFFERENT STATE-OF-THE-ART SINGLE SYSTEMS AND BASELINE SYSTEMS

Model	EER (%)
Melspectrogram-LCNN with DA [26]	16.96
Our proposed model with bi-level	**20.24**
Melspectrogram-LCNN w/o DA [26]	21.90
B04: RawNet2 [6]	22.38
B03: LFCC-LCNN [3]	23.48
RawGAT [7]	23.71
Res2Net [27]	24.47
B02: LFCC-GMM [3]	25.25
B01: CQCC-GMM [3]	25.56
Resnet18-AM-Softmax [20]	28.81
Resnet18-OC-Softmax [20]	31.44

D. Ablation study

We conduct a series of ablation experiments, which is an intuitive way to demonstrate the effectiveness of the design choices for our proposed model.

Impact of bi-level optimization. In order to show the effectiveness of the bi-level optimization technique, we compare the detection performance of our model both with and without bi-level optimization. For both versions, with or without bi-level optimization, the models are trained with the same training dataset composed of known attacks D_k and unknown attacks D_u. The remaining structure and the hyperparameter settings also stay the same. The results are shown in Table II. We can observe that, without the bi-level optimization, the performance on the DF evaluation set degrades significantly by 43% in terms of EER (35.54% cf. 20.24%). It indicates that the original Transformer-based model suffers from overfitting due to insufficient training data.

TABLE II
PERFORMANCE IMPACT DUE TO BI-LEVEL OPTIMIZATION

Model	Validation Loss	EER on Evaluation Set (%)
w/o bi-level	0.328	35.54
w/ bi-level	**0.321**	**20.24**

Impact of parameter update frequency during training. As mentioned in Section 3.4, we adopt a local optimal value of θ_2 to optimize θ_1 by updating θ_m^B to θ_m^A. However, the frequency of parameter updating may affect the detection performance. We compare the performance of models trained with different update frequencies. The approach that we adopt in the final proposed version, as Algorithm 1 states, is that θ_m^A is updated by θ_m^B every epoch. We also try to update θ_m^A more frequently, such as once per mini-batch or once per hundred times of mini-batches. As Table III shows, the performance decreases as the update frequency increases. It is because, with more iterations of training, θ_m^B can better reach a local optimal value to describe the characteristics on D_u set. Also, it should be noted that all three versions of bi-level optimization with different parameter update frequencies outperform the system without bi-level optimization.

TABLE III
PERFORMANCE IMPACT DUE TO VARIOUS PARAMETER UPDATE FREQUENCY FOR BI-LEVEL OPTIMIZATION

Model	EER (%)
Update once per mini-batch	24.01
Update once per 100 mini-batches	21.74
Update once per epoch **(Used)**	**20.24**

We also perform experiments to evaluate the impact of the Transformer structure on the detection result. We find that a larger number of encoder blocks and attention heads does not necessarily improve the evaluation performance because of overfitting. The selection of six encoders with 4 heads gives the best detection result on the evaluation set. Using ImageNet-pre-trained weights [28] in Transformer architecture will also lead to overfitting.

V. CONCLUSION

In this paper, we present a novel Deepfake audio detection model with great improvement in robustness and generalization ability. We propose a Transformer-based classifier trained on both learned and hand-crafted features from the raw audio inputs. Combining these two forms of features produces a more generalizable representation of Deepfake audios with varying characteristics. We formulate the training phase as a bi-level problem to optimize the parameters to detect unseen spoofing attacks. Our proposed model achieves the best detection performance on the ASVspoof2021 DF track as a single model without model fusion and data augmentation, which outperforms the best official baseline with a 9.56% improvement in the EER metric. We also conduct systematic ablation studies to show the effectiveness of bi-level optimization to gain the robustness of spoofing detection models and contribute to reducing overfitting.

REFERENCES

[1] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[2] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[3] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. ASVspoof Challenge workshop*, 2021, pp. 47–54.

[4] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[5] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with SincNet," *arXiv preprint arXiv:1811.09725*, 2018.

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

[7] H. Tak, J. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[8] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *Proc. Interspeech 2019*, pp. 1078–1082, 2019.

[9] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[10] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1018–1025.

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech 2015*, 2015, pp. 2087–2091.

[12] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 35–41.

[13] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, pp. 235–256, 2007.

[14] A. Sinha, P. Malo, and K. Deb, "Efficient evolutionary algorithm for single-objective bilevel optimization," *arXiv preprint arXiv:1303.3901*, 2013.

[15] M. Xu, W. Wang, S. Luo, C. Shi, Y. Bengio, R. Gomez-Bombarelli, and J. Tang, "An end-to-end framework for molecular conformation generation via bilevel programming," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 537–11 547.

[16] W. Yang, R. Cai, and A. Kot, "Image inpainting detection via enriched attentive pattern with near original image augmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2816–2824.

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

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

[19] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop labs' submission to the ASVspoof 2021 challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[20] 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.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[23] E. R. Bartusiak and E. J. Delp, "Synthesized speech detection using convolutional transformer-based spectrogram analysis," *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021.

[24] C. Li, F. Yang, and J. Yang, "The role of long-term dependency in synthetic speech detection," *IEEE Signal Processing Letters*, vol. 29, p. 1142–1146, 2022.

[25] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "ASVspoof 2021 challenge - speech deepfake database," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4835108

[26] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC antispoofing systems for the ASVspoof2021 challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[27] 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)*. IEEE, 2021, pp. 6354–6358.

[28] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.