



Backend Ensemble for Speaker Verification and Spoofing Countermeasure

Li Zhang, Yue Li, Huan Zhao, Qing Wang, Lei Xie*

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University (NPU), Xi'an, China

lizhang.aslp.npu@gmail.com, lxie@nwpu.edu.cn

Abstract

This paper describes the NPU system submitted to Spoofing Aware Speaker Verification Challenge 2022. We particularly focus on the *backend ensemble* for speaker verification and spoofing countermeasure from three aspects. Firstly, besides simple concatenation, we propose circulant matrix transformation and stacking for speaker embeddings and countermeasure embeddings. With the stacking operation of newly-defined circulant embeddings, we almost explore all the possible interactions between speaker embeddings and countermeasure embeddings. Secondly, we attempt different convolution neural networks to selectively fuse the embeddings' salient regions into channels with convolution kernels. Finally, we design parallel attention in 1D convolution neural networks to learn the global correlation in channel dimensions as well as to learn the important parts in feature dimensions. Meanwhile, we embed squeeze-and-excitation attention in 2D convolutional neural networks to learn the global dependence among speaker embeddings and countermeasure embeddings. Experimental results demonstrate that all the above methods are effective. After fusion of four well-trained models enhanced by the mentioned methods, the best SASV-EER, SPF-EER and SV-EER we achieve are 0.559%, 0.354% and 0.857% on the evaluation set respectively. Together with the above contributions, our submission system achieves the fifth place in this challenge.

Index Terms: speaker verification, spoofing countermeasure, backend ensemble

1. Introduction

The key of automatic speaker verification (ASV) is to decide whether two speech utterances come from the same speaker [1, 2]. With the development of deep neural network (DNN) and easy availability of computing resources and massive data, ASV technology has delivered the high accuracy required in voice-enabled IoT gadgets control, speech authorization and forensic applications [3, 4, 5]. However, ASV systems are vulnerable under various kinds of malicious spoofing attacks, i.e., specially crafted utterances generated by adversaries to deceive the ASV system and to provoke false accepts [6, 7, 8, 9].

The data scenarios for speaker spoofing include logical access (LA), physical access (PA), speech deep fake (DF) [10]. The generated human-like speech deceiving ASV systems poses a great threat to the security of society if misused malignantly. Fortunately, this problem has attracted the attention of many researchers. There are several anti-spoofing challenges [11, 12] held to boost the development of countermeasure (CM) systems to help detecting spoofing attacks [13, 14, 15, 16]. While most CM systems achieve fabulous performance in detecting spoofing speech, they seriously affect the performance of the

zero-effort impostors' detection when they work with ASV systems [17]. Spoofing Aware Speaker Verification Challenge 2022 (SASV) [6] provides a common platform for researchers to extend the focus of ASVspoof upon CMs to the consideration of integrated systems where both CM and ASV subsystems are optimized jointly to improve reliability [6]. The SASV challenge focuses on spoofing attacks generated using speech synthesis/text-to-speech (TTS) and voice conversion (VC), which are LA spoofing. It proposes to jointly optimize the automatic ASV system and spoofing CM system, which aims to develop a new spoofing-aware speaker verification system.

Recently, several works pay attention to the ensemble of ASV and CM systems considering both performances of the two systems. The ensemble methods can be classified into three types, which are the tandem for ASV and CM systems to make a decision of speaker verification, multi-task learning frameworks for ASV and CM systems, backend ensemble for ASV and CM systems. In terms of the tandem for ASV and CM systems, Tomi et al. [17] proposed several new tandem detection cost function (t-DCF) to assess the reliability of spoofing CMs deployed in tandem with an ASV system. Kanervisto et al. [18] optimized the tandem system directly by creating a differentiable version of t-DCF and employing techniques for reinforcement learning. For the multi-task learning frameworks for ASV and CM systems, Li et al. [19] proposed a multi-task learning neural network to make a joint system of ASV and anti-spoofing, which verified that joint optimization was more advantageous than the cascaded systems. Moreover, Li et al. [20] proposed a multi-task learning framework with contrastive loss to joint decision of anti-spoofing and ASV. As for the backend ensemble for ASV and CM systems, Gomez-Alanis et al. [21] developed an integration neural network and a loss function based on the minimization of the area under the expected (AUE) performance and spoofability curve (EPSC) to jointly process the embeddings extracted from ASV and anti-spoofing systems. Chettri et al. [22] combined deep neural networks and traditional machine learning models as ensemble models by logistic regression to integrate spoofing detection in an ASV system.

In this challenge, we focus on exploring the backend embedding ensemble of ASV and CM systems. Our target is to build a backend ensemble system trained with speaker embeddings and CM embeddings to make a joint decision of speaker verification. The primary work of the backend ensemble system is to mine as much effective speaker discrimination information and spoofing detection information as possible from speaker embeddings and CM embeddings. Specifically, we exploit different embedding fusion methods, neural network frameworks and enhanced attention mechanisms to aggregate valuable information for spoofing aware speaker verification. Together with the proposed methods, our experimental results on the evaluation set have significant improvements compared with the baseline results in the challenge.

* Corresponding author.

2. System Overview

Figure 1 gives an overview of our system. It consists of three modules which are embedding extractors (ECAPA-TDNN and AASIST), embedding fusion and ensemble model. The outputs are labels of test trials denoting whether the enrollment and test utterances belong to the same speaker. Our contributions are to explore different embedding fusion methods, ensemble network frameworks and attention mechanisms in the ensemble modules to learn shared speaker and detection information from ASV and CM systems for spoofing aware speaker verification.

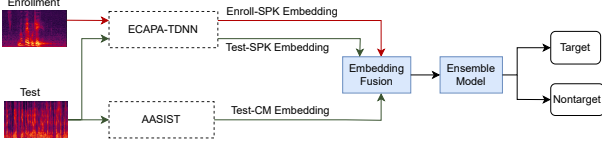


Figure 1: The overview of backend ensemble system.

In the embedding extractor module, the speaker embedding extractor and CM embedding extractor we use are pretrained ECAPA-TDNN [3] and AASIST [23] offered by the challenge organizer [6]. Therefore, the two embedding extractors in the dashed boxes do not participate in training in our system.

In the embedding fusion module, we exploit three kinds of embedding fusion methods which are concatenation, circulant matrix transformation and stacking.

For the ensemble model module illustrated in Figure 2, besides the enlarged DNN baseline, we introduce another two backend frameworks based on convolutional neural networks (CNN). Firstly, we enlarge the baseline2 model which is a 3-layer DNN [6], but the performance saturates as the model gets deeper. Then we stack the embeddings and use 1D CNN to derive the decision of speaker verification. Meanwhile, we propose parallel attention after convolutional blocks to learn the global relationship in channel dimensions as well as to learn the important regions in feature dimensions. Finally, we propose to make circulant matrix transformation on embeddings and stack newly-defined embeddings before feeding them into a 2D CNN with squeeze-and-excitation (SE) attention [24, 25].

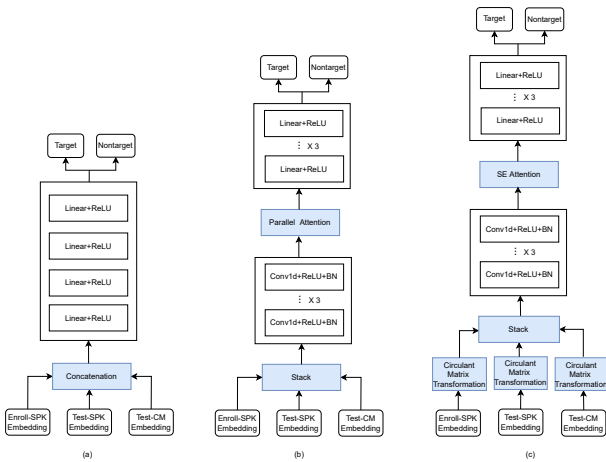


Figure 2: Different backend ensemble modules: (a) is the enlarged baseline model (DNN), (b) is the 1D CNN with parallel attention, and (c) is the circulant matrix transformation for 2D CNN with SE attention.

3. Backend Ensemble Frameworks

The performance of the baseline2 model provided by challenge organizer [6] is easy to saturate as the model gets deeper. Therefore, we propose another two CNN-based backend ensemble frameworks for speaker verification and spoofing countermeasure.

3.1. 1D CNN with Parallel Attention

We adopt 1D CNN to process the stacked speaker embeddings and CM embeddings. The aim is to use convolutional kernels for selectively fusing the different embedding regions into channels. Then we introduce a parallel attention (PA) module to learn global interactions among different embeddings and focus on important areas in the feature dimension, as illustrated in Figure 3. The PA module learns attention masks along with the channel dimension and feature dimension respectively.

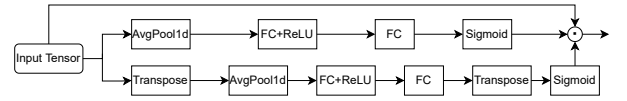


Figure 3: The parallel attention (PA) in 1D CNN.

Suppose the input tensor generated from 1D convolutional blocks is $T^{(B \times C \times F)}$, where B, C, F are batchsize, channel dimension and feature dimension respectively. We split two paths to calculate the channel-wise and feature-wise attention at the same time. The average pooling on the feature dimension and channel dimension of parallel attention is calculated as:

$$T^{B \times F} = \frac{1}{C} \sum_{i=1}^C T^{(B \times C \times F)}, \quad (1)$$

$$T^{B \times C} = \frac{1}{F} \sum_{i=1}^F \tau(T^{(B \times C \times F)}), \quad (2)$$

where τ in Eq. 2 is to transpose the channel dimension and feature dimension in $T^{B \times C \times F}$. Then we use linear connection layers to learn the channel-wise and feature-wise attention in parallel. The equations are:

$$T_1^{B \times F} = \rho((T^{B \times F} \otimes W_1^{F \times F/r}) \otimes W_2^{F/r \times F}), \quad (3)$$

$$T_2^{B \times C} = \rho((T^{B \times C} \otimes W_3^{C \times C/r}) \otimes W_4^{C/r \times C}), \quad (4)$$

where \otimes is matrix multiplication, r is the middle layer dimension reduction factor and ρ is the *sigmoid* operation. The weighted intermediate tensor $T^{B \times C \times F}$ is calculated as:

$$(T^{B \times C \times F})' = T_2^{B \times C} \cdot T^{B \times C \times F} \cdot T_1^{B \times F}, \quad (5)$$

where \cdot is dot products with automatic dimension expansion. $(T^{B \times C \times F})'$ is the output tensor of the PA module.

3.2. Circulant Matrix Transformation for 2D CNN with SE Attention

To further use convolutional kernels for selectively fusing the embeddings' salient regions, we adopt circulant matrix transformation to transfer speaker embeddings and CM embeddings into circulant matrices. The circulant matrix is a square matrix,

in which all row vectors are composed of the same elements and each row vector is rotated one element to the right relative to the preceding row vector [26]. Suppose the enrollment speaker embedding, test speaker embedding and test CM embedding are $e_i = \{x_1, x_2, x_3 \dots x_d\}_{1:d}$, $t_i = \{s_1, s_2, s_3 \dots s_b\}_{1:b}$ and $c_i = \{m_1, m_2, m_3 \dots m_q\}_{1:q}$, the circulant matrix transformation are formulated as follows.

$$e_i' = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_d \\ x_d & x_0 & x_1 & x_2 & \dots \\ x_{d-1} & x_d & x_1 & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & \dots & x_d & x_0 \end{pmatrix}, \quad (6)$$

$$t_i' = \begin{pmatrix} s_1 & s_2 & s_3 & \dots & s_b \\ s_b & s_0 & s_1 & s_2 & \dots \\ s_{b-1} & s_b & s_1 & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_1 & s_2 & \dots & s_b & s_0 \end{pmatrix}, \quad (7)$$

$$c_i' = \begin{pmatrix} m_1 & m_2 & m_3 & \dots & m_q \\ m_q & m_0 & m_1 & m_2 & \dots \\ m_{q-1} & m_q & m_1 & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_1 & m_2 & \dots & m_q & m_0 \end{pmatrix}, \quad (8)$$

where b, d, q are the embedding dimensions. As each row of the circulant matrix shifts one element, with newly-defined interaction operations, we almost explore all possible interactions among different embeddings. After that, we stack the three circulant matrices into one tensor and feed it into a 2D CNN. Meanwhile, we embed the SE attention [24] after the last convolutional layer to learn the dependency of different embeddings.

4. Experimental Setup

4.1. Datasets

All the experiments presented further are conducted on the logical access (LA) training, development and evaluation sets in ASVspoof 2019 database [27]. The assessment results are derived from ASV trials in development and evaluation sets [6].

4.2. Model Configurations

In this paper, eight models are trained with the following configurations.

- **Extend(512)_DNN** The neural nodes of 4-layer DNN are [512, 256, 128, 64].
- **Extend(1024)_DNN** The neural nodes of 5-layer DNN are [1024, 512, 256, 128, 64].
- **1D.CNN** There are 3-layer 1D convolution, 1-layer adaptive average pooling and 3-layer DNN in the 1D.CNN model. The channels and kernels of 3-layer 1D convolution are [256, 128, 64] and [3, 3, 3]. Each convolutional layer is followed with a normalization layer and a LeakReLU activation layer. The neural nodes of 3-layer DNN are [512, 256, 64].
- **1D.CNN_SE** The SE attention layer is embedded after the third convolutional network layer in 1D.CNN. The inner channel reduction ratio of SE attention is eight. Other configurations of this model are the same as those in the 1D.CNN model.
- **1D.CNN_PA** The parallel attention layer is embedded after the third convolutional network layer in 1D.CNN. The inner channel reduction ratio of parallel attention is eight. Other configurations of this model are the same as those in the 1D.CNN model.

- **2D.CNN** There are 4-layer 2D convolution, 1-layer adaptive average pooling and 3-layer DNN in the 2D.CNN model. The channel and kernel configures are [32, 64, 128, 256] and [5, 3, 3, 3]. The configuration of adaptive average pooling is [16, 16]. The neural nodes of DNN are [256, 128, 64].

- **2D.CNN_SE** The SE attention [24] layer is embedded after the third convolutional network layer. Other configurations of this model are the same as those in the 2D.CNN model.

- **2D.CNN_VSE** In addition to SE attention, we also use a variant of SE attention [25] in 2D.CNN model. The variant SE attention (VSE) has been demonstrated to be effective in speaker verification [28]. Other configurations of this model are the same as those in the 2D.CNN model.

4.3. Training Setup

We adopt the well pre-trained ECAPA-TDNN [3] and AASIST [23] models provided by the challenge organizer [6] as our embedding extractors. In the training step, the embedding extractors are fixed without joint training. The initial learning rate is 1e-3. The optimizer is Adam with keras scheduler and weight decay is 1e-3. The loss function is cross entropy with bias-weight [0.1, 0.9] because of the unbalance of bonafide and spoofing datasets in ASVspoof 2019 train set.

4.4. Score Metric

The classical equal error rate (EER) [29] is used as the primary metric. Specifically, there are three kinds of EER metrics, i.e., SASV-EER, SPF-EER and SV-EER [30]. The SASV-EER does not distinguish between different speaker (zero-effort, non-target, or impostor) access attempts and spoofed access attempts. SPF-EER is the metric of spoofing attacks and target trials. SV-EER is the metric of traditional ASV trials without spoofing attacks [6].

5. Experimental Results

Experimental results in Table 1 are derived from the models trained with ASVspoof 2019 training set. We can see that the SASV-EER of Extend (512)_DNN model has an absolute reduction of 1.529% compared with that of baseline2 model [6] on the evaluation set. After we stack speaker embeddings and CM embeddings and feed them into the 1D.CNN model, the SASV-EER on the evaluation set dramatically reduces to 1.361%. The performance improvements are thanks to the decline in SV-EER. But the SPF-EER of 1D.CNN on evaluation set slightly rises. When we embed SE attention into 1D.CNN model, the SPF-EER gets more reduction. Meanwhile, the SASV-EER of the evaluation set is reduced by 0.244% compared with that of 1D.CNN. After we replace the SE attention with the parallel attention in 1D.CNN_SE, the SASV-EER on the evaluation set is further reduced by 0.093%. With the help of circulant matrix transformation, the SASV-EER of 2D.CNN model is better than that of 1D.CNN model on the evaluation set. With adding SE attention in 2D.CNN, the SASV-EER of 2D.CNN_SE model even reaches 0.998% on the evaluation set, and we get the lowest SPF-EER and SV-EER compared with other models in Table 1. Moreover, when we embed the VSE attention in 2D.CNN model, the SASV-EER of the evaluation set is decreased by 0.134% compared with that of 2D.CNN model.

Table 1: EER (%) results (trained with training set) on SASV 2022 development and evaluation partitions.

| Model Index | Model Name | DEV | | | EVAL | | |
|-------------|-------------------|----------|---------|--------|--------------|--------------|--------|
| | | SASV-EER | SPF-EER | SV-EER | SASV-EER | SPF-EER | SV-EER |
| A | Model_baseline2 | 4.85 | 0.13 | 12.87 | 6.37 | 0.78 | 11.48 |
| B | Extend (512)_DNN | 3.973 | 0.193 | 9.097 | 4.841 | 0.797 | 8.429 |
| C | Extend (1024)_DNN | 3.705 | 0.1634 | 9.652 | 4.926 | 0.710 | 8.683 |
| D | 1D_CNN | 0.606 | 0.135 | 1.456 | 1.361 | 1.135 | 1.750 |
| E | 1D_CNN_SE | 0.876 | 0.110 | 1.699 | 1.117 | 0.519 | 1.638 |
| F | 1D_CNN_PA | 1.022 | 0.103 | 1.954 | 1.024 | 0.819 | 1.378 |
| G | 2D_CNN | 0.687 | 0.067 | 1.752 | 1.212 | 0.416 | 2.019 |
| H | 2D_CNN_SE | 0.846 | 0.135 | 2.167 | 0.998 | 0.497 | 1.582 |
| I | 2D_CNN_VSE | 1.257 | 0.134 | 2.409 | 1.078 | 0.509 | 1.765 |

Table 2: EER (%) results (trained with training and dev sets) on SASV 2022 development and evaluation partitions.

| Model Index | Model Name | DEV | | | EVAL | | |
|-------------|------------------|----------|---------|--------|--------------|--------------|--------------|
| | | SASV-EER | SPF-EER | SV-EER | SASV-EER | SPF-EER | SV-EER |
| B_Aug | Extend (512)_DNN | 0.011 | 8.97e-5 | 0.017 | 3.026 | 0.837 | 5.497 |
| D_Aug | 1D_CNN | 0.011 | 0.067 | 1.666 | 0.837 | 0.350 | 1.303 |
| E_Aug | 1D_CNN_SE | 0.078 | 8.97e-5 | 0.134 | 0.812 | 0.570 | 0.981 |
| F_Aug | 1D_CNN_PA | 0.067 | 0.066 | 0.022 | 0.768 | 0.465 | 1.053 |
| G_Aug | 2D_CNN | 0.122 | 0.033 | 0.202 | 0.760 | 0.346 | 1.224 |
| H_Aug | 2D_CNN_SE | 0.078 | 0.067 | 0.202 | 0.758 | 0.476 | 1.125 |
| I_Aug | 2D_CNN_VSE | 0.134 | 0.002 | 0.269 | 0.734 | 0.416 | 1.104 |

Table 3: Fusion EER (%) results on SASV 2022 development and evaluation partitions

| Model Index | Fusion Method | DEV | | | EVAL | | |
|-----------------|-------------------|----------|---------|--------|--------------|--------------|--------------|
| | | SASV-EER | SPF-EER | SV-EER | SASV-EER | SPF-EER | SV-EER |
| D_Aug & G_Aug & | Linear Regression | - | - | - | 0.838 | 0.838 | 1.136 |
| H_Aug & I_Aug | Score Averaging | 0.067 | 0.067 | 0.135 | 0.559 | 0.354 | 0.857 |

To further improve the performance of the proposed systems, we add the ASVspoof 2019 development set to train the above models in Table 1. The results are illustrated in Table 2. With the help of extra data in training, the SASV-EER of Extend(512)_DNN model declines absolute 1.815%. Moreover, the SASV-EERs of all the models are less than 0.998% on the evaluation set except that of Extend(512)_DNN model. At the same time, we find both the SPF-EER and SV-EER have further reduction. The best single model in this paper is the 2D_CNN_VSE, which is trained with the training and development sets of ASVspoof 2019. The SASV-EER, SPF-EER and SV-EER of the best model are 0.734%, 0.416% and 1.104% respectively. The results demonstrate the proposed embedding fusion methods, ensemble model structures and attention mechanisms have significant contributions to the backend ensemble of speaker verification and spoofing countermeasure systems.

Finally, we fuse the above models trained with the ASVspoof 2019 training and development set. We adopt two fusion methods, i.e., linear regression in Bosaris toolkit [31] and score averaging. The fusion results are shown in Table 3. After averaging the scores of the 1D_CNN model and three 2D CNN-based models in Table 2, the SASV-EER, SPF-EER and SV-EER reach 0.559%, 0.354% and 0.857% respectively. We notice that the fusion of linear regression leads to worse results

compared with those of score averaging. This is probably because the development set is used to train the models, which leads the models to overfit to the development set. Meanwhile, the experimental results on the development set from Table 1 and Table 2 illustrate the models overfit to the development set when we add development set to train the models.

6. Conclusions

This paper develops a spoofing aware speaker verification system for the SASV challenge. In this challenge, we particularly focus on exploring the backend ensemble schemes from three aspects, which are embedding fusion methods, different ensemble frameworks and attention mechanism design respectively. Specifically, we propose circulant matrix transformation and stacking operation for embedding fusion. With the fused embeddings, we further introduce different backend ensemble frameworks based on CNN. Finally, we introduced parallel attention and SE attention in CNN-based frameworks to capture the global relationship between speaker embeddings and countermeasure embeddings. With the proposed method, the SASV-EER of our best single backend ensemble system reaches 0.734% on the evaluation set. Together with the above contributions, the fusion on four well-trained models in this paper finally leads our submission to the fifth place in the SASV challenge.

7. References

- [1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, pp. 1–22, 2004.
- [2] R. Naika, "An overview of automatic speaker verification system," *Intelligent Computing and Information and Communication*, pp. 603–610, 2018.
- [3] B. Desplanques, J. Thienpondt, and K. Demuynck, "PECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech*, 2020.
- [4] J. P. Giraldo, S. Lauwereins, K. Badami, H. Van Hamme, and M. Verhelst, "18 μ w soc for near-microphone keyword spotting and speaker verification," in *2019 Symposium on VLSI Circuits*. IEEE, 2019, pp. C52–C53.
- [5] T. J. Machado, J. Vieira Filho, and M. A. de Oliveira, "Forensic speaker verification using ordinary least squares," *Sensors*, vol. 19, no. 20, p. 4385, 2019.
- [6] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.
- [7] 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.
- [8] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: review and analysis," *International Journal of Speech Technology*, pp. 1–30, 2021.
- [9] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *Proc. Interspeech*, 2019.
- [10] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *ASVspoof 2021 Workshop Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.
- [11] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *ASVspoof 2021 Workshop Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.
- [12] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "ADD 2022: the first audio deep synthesis detection challenge," *ADD ICASSP workshop*, 2022.
- [13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP*. IEEE, 2021, pp. 6369–6373.
- [14] C. Hanilçi, "Linear prediction residual features for automatic speaker verification anti-spoofing," *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 16 099–16 111, 2018.
- [15] A. Chadha, A. Abdullah, L. Angeline, and S. Sivanesan, "A review on state-of-the-art automatic speaker verification system from spoofing and anti-spoofing perspective," *Indian Journal of Science and Technology*, vol. 14, no. 40, pp. 3026–3050, 2021.
- [16] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning," in *Interspeech*, 2019, pp. 1048–1052.
- [17] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.
- [18] A. Kanervisto, V. Hautamäki, T. Kinnunen, and J. Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 477–488, 2021.
- [19] J. Li, M. Sun, and X. Zhang, "Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1517–1522.
- [20] J. Li, M. Sun, X. Zhang, and Y. Wang, "Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss," *IEEE Access*, vol. 8, pp. 7907–7915, 2020.
- [21] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2020.
- [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," *Proc. Interspeech*, 2019.
- [23] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *Proc. ICASSP*, 2022.
- [24] 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.
- [25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.
- [26] R. M. Gray, "Toeplitz and circulant matrices: A review," 2006.
- [27] 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," in *INTERSPEECH 2019*, Graz, Austria, Sep. 2019.
- [28] L. Zhang, Q. Wang, and L. Xie, "Duality temporal-channel-frequency attention enhanced speaker representation learning," *ASRU*, 2021.
- [29] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.
- [30] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, and T. Kinnunen, "Sasv 2022: The first spoofing-aware speaker verification challenge," *arXiv preprint arXiv:2203.14732*, 2022.
- [31] N. Brümmer and E. De Villiers, "The Bosaris toolkit: Theory, algorithms and code for surviving the new DCF," *arXiv preprint arXiv:1304.2865*, 2013.