Explore Backend Ensemble of Speaker Verification and Spoofing Countermeasure

Li Zhang, Yue Li, Huan Zhao, 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

Abstract

This paper describes the backend ensemble system of speaker verification and spoofing countermeasure submitted to Spoofing Aware Speaker Verification Challenge 2022. The primary work of the backend ensemble system is mining as much as effective speaker and detecting characteristics from speaker embeddings and countermeasure embeddings. We exploit different embedding mixture methods, neural network frameworks and enhanced attention mechanisms to aggregate valuable information for speaker verification. Specifically, we propose to stack and transfer speaker embeddings and countermeasure embeddings into circulant matrices to facilitate using convolutional kernels selectively fusing the embeddings' salient regions into channel-wise. Meanwhile, we design a global attention module to capture interactions of different embeddings. With the proposed global attention mechanism, the SASV EER, SPF EER and SV EER of 2D convolutional neural network with the global attention are dramatically decreased by 2.258%, 0.372%, 4.444%. In addition, with the proposed circulant embedding matrices, the SASV EER, SPF EER and SV EER of 2D convolutional neural network with variant squeeze-and-excitation attention reach 0.734%, 0.416% and 1.104%. After fusion of four well-trained models in the paper, the best SASV EER, SPF EER and SV EER we achieve are 0.559%, 0.354% and 0.857% on the evaluation set.

Index Terms: speaker verification, spoofing countermeasure, backend ensemble

1. Introduction

Automatic Speaker Verification (ASV) attempts to decide whether a pair of speech is from the same speaker [1]. 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 [2, 3, 4]. 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 [5, 6, 7, 8].

The data scenarios for speaker spoofing include logical access (LA), physical access (PA), speech deep fake (DF) [9]. The generated human-like speech deceiving ASV systems poses a great threat to the security of society if misused malignantly. Fortunately, this problem has intrigued the attention of many researchers. There are many anti-spoofing challenges [10, 11] held to boost the development of countermeasure systems (CM) to help detecting spoofing attacks [12, 13, 14, 15]. 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 [16]. SASV [5] 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 [5]. SASV challenge focuses on spoofing attacks generated using speech synthesis/text-to-speech (TTS), 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, the ensemble works of ASV and CM systems considering both performances have sprouted thanks to the driven of challenges like SASV and ASVspoofing challenges [5] in the speech community. Li et al. [17] 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 cascaded systems with traditional methods. Kanervisto et al. [18] optimized the tandem system directly by creating a differentiable version of t-DCF and employing techniques from reinforcement learning. Chettri et al. [19] combined both deep neural networks and traditional machine learning models as ensemble models through logistic regression to integrate spoofing detection in a ASV system. Li et al. [20] proposed a multi-task learning framework with contrastive loss to joint decision of anti-spoofing and ASV. 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. Zhang et al. [22] proposed a probabilistic framework for fusing the ASV and CM subsystem scores.

In this challenge, we focus on exploring the backend embedding ensemble of ASV and countermeasure 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. We attempt different embedding mixture methods, ensemble model frameworks and attention mechanisms to dig effective characteristics from speaker embeddings and CM embeddings for speaker verification. The well-experimental results demonstrate the effectiveness of our proposed global attention and the circulant matrices transformation methods.

The rest of this paper is organized as follows: Section 2 describes the overview of our system. Section 3 details different backend ensemble frameworks. Section 4 presents the experimental setup of this paper. Section 5 elaborates the experimental results. Finally, section 6 concludes this paper.

[*] Corresponding author.

2. System Overview

The system overview in this paper is illustrated in Figure 1. It consists of embedding extractors and backend ensemble modules. The outputs are labels of test trials denoting whether the enrollment and test utterances belong to the same speaker. The speaker embedding extractor and CM embedding extractor we use are well pretrained ECAPA-TDNN [2] and AASIST [23] offered by official organizers [5]. The two embedding extractors in the dotted boxes are not participating in training. Our contributions are to explore different embedding mixture methods, ensemble network frameworks and attention mechanisms of the backend ensemble module to dig more effective information for speakers' certification.

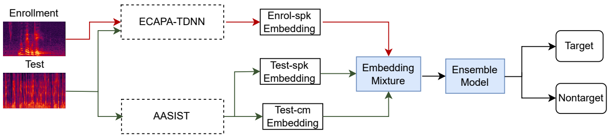

Figure 1: *The overview of backend ensemble system.*

We exploit three kinds of backend ensemble modules illustrated in Figure 2. Firstly, we enlarge the baseline2 model which is a 3-layer DNN [5], but the performance saturates as the model gets deeper. We believe that mining the interactions of speaker embeddings and CM embeddings facilitate the correct decision of speaker authentication. Then we stack the above three embeddings and use 1D convolution neural work (1D CNN) to derive the decision of speaker verification. Meanwhile, we propose a global attention submodule after convolutional blocks to learn the independent information among embeddings. Finally, we propose to use circulant matrix transformation to derive two-dimension of embedding from original one-dimension embeddings and then feed the stacked three two-dimensions embeddings into a 2D convolution neural network (2D CNN) with squeeze-and-excitation attention (SEA) [24]. The specific design of each module will be explained in the next section.

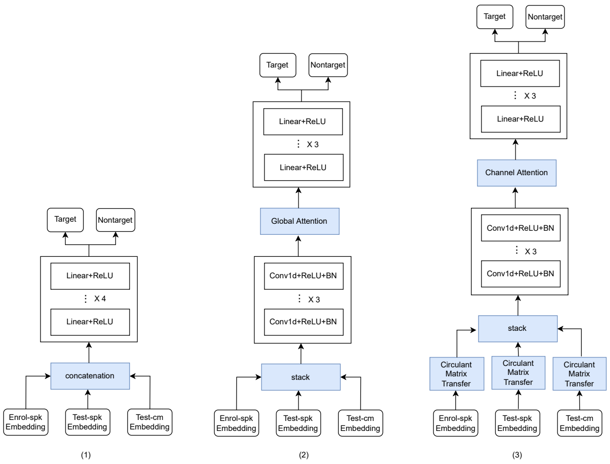

Figure 2: *Different backend ensemble modules. (1) is the enlarged baseline2 model. (2) is the 1D CNN with global attention. (3) is the circulant matirx transformation in 2D CNN with channel attention*

3. Backend Ensemble Frameworks

3.1. Extend Baseline Model

The baseline2 model provided by challenge officials [5] is a 3-layer DNN model with neural node configuration [256,128,64]. We extend the node configuration into [512, 256, 128, 64] and [1024, 512, 256, 128] which are expressed as Eextend (512)_DNN and Extend(1024)_DNN respectively.

3.2. 1D CNN with Global Attention

Next we adopt 1D CNN to process the stacked three kinds of embeddings. The motivation is to use convolutional kernels selectively fusing the different embedding information into channel-wise. Then we propose a global attention module to learn global context information in channel-wise and feature-wise. Specifically, the global attention learns attention masks along with the channel-wise and feature-wise respectively, which aims to learn the interdependence of three embeddings and pure the useful information for speaker verification. The global attention is illustrated in Figure 3.

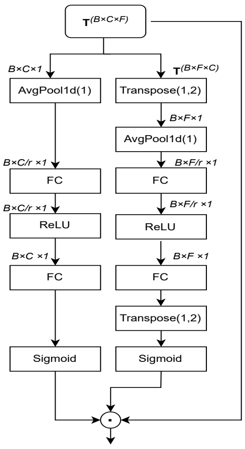

Figure 3: *Global attention (GA).*

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

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

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

where τ in 2 is to transpose the channel-wise and feature-wise in $T^{B \times C \times F}$.

Then we use linear connection layers to learn the channel-wise and feature-wise attention in parallel.

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

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

where 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 formula:

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

where · are dot products with automatic dimension expansion.

3.3. 2D Convolutional Neural Network with SE attention

In order to learn the interactions among enrollment speaker embeddings, test speaker embeddings and CM embeddings, we transfer the above three embeddings into circulant matrices. Circulant matrices are square matrices 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 [25]. Suppose the enrollment speaker embedding, test speaker embedding and test CM embedding are $e_i = \{x_1, x_2, x_3 ... x_d\}_{1:d}$, $t_i = \{s_1, s_2, s_3 ... s_b\}_{1:b}$ and $c_i =$

$\{m_1, m_2, m_3 ... m_q\}_{1:q}$, the circulant matrix transformation are formulated as following:

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

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

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

As each row of the circulant matrix shifts one element, with newly-defined interaction operations, we almost explore all possible interactions between different embeddings. After that, we stack the three circulant matrices into one tensor and feed it into a 2D convolutional neural network. Meanwhile, we embed channel-wise attention [24] behind the last convolutional layer to learn the dependency of different embeddings.

4. Experimental Setup

4.1. Datasets

All experiments presented further are conducted on ASVspoof 2019 logical access (LA) train and development partitions [6]. The development and evaluation protocols are ASVspoof 2019.LA.asv.eval.gi.trl.txt and ASVspoof 2019.LA.asv.eval.gi.trl.txt from ASVspoof 2019 challenge [6]. We have attempted to do data augmentation on raw audio with MUSAN [26] and Room Impulse Response and Noise (RIRs) databases [27], but there is no improvement in the experimental results. In the future, we will further research on the data augmentation method in this topic.

4.2. Model Configurations

We have trained six models in total. The configurations of them are in following:

- **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. 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_SEA** The SE attention layer is embedded after the third convolutional network layer. The inner channel reduction ratio of SE attention is eight. The other configurations of CNN model are the same as 1D_CNN model's.
- **1D_CNN_GA** The global attention layer is embeded after the third convolutional network layer. The inner channel reduction ratio of global attention is eight. The other configurations of CNN model are the same as 1D_CNN model's.
- **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_SEA** The SE attention layer is embedded after the third convolutional network layer. The inner channel reduction ratio of SE attention is eight. The other configurations of CNN model are the same as 2D_CNN model's.
- **2D_CNN_VSE** In addition to SE attention, we also tried a variant of SE attention [28] which has been demonstrated to be effective in improving the performance of speaker verification system [29].

4.3. Training Setup

We adopt the well pre-trained ECAPA-TDNN [2] and AASIST [23] models provided by the challenge official [5] as our embedding extractors. In training step, the embedding extractors are fixed without joint training. The batchsize is 1024 and initial learning rate is 1e-3. The optimizer is Adam with keras scheduler and weigt decay 1e-3. The loss function is cross entropy with bias-weight [0.1, 0.9] because of the inbalance of bonafide and spoofing datasets in ASVspoof 2019 train set.

4.4. Score Metric

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

5. Experimental Results

Experimental results in Table 1 are derived from the models trained with ASVspoof 2019 train set partition. We can see that the SASV EER of Extend (512)_DNN model achieves 1.529% absolute reductions compared with baseline2 model in [5] on ASVspoof 2019 evaluation set. After we stack speaker embeddings and CM embedding and feed them into a 1D CNN model, the SASV EER on SASVspoof 2019 evaluation set dramatically reduces into 1.361%. The main reason for the performance improvement thanks to the reductions of SV EER. But the SPF EER has slight rise. When we embeded SE attention into 1D CNN model, the SPF-EER reduces to lower. Meanwhile, the SVSA EER of evaluation set has 0.224% reductions compared with that of 1D CNN. We replace the SE attention with our proposed global attention, the SASV EER of evaluation set has further 0.093% absolute reductions. The SASV EER of 2D CNN with circulant matrix transformation is better than that of 1D CNN model on evaluation set. Then we add SE attention module in 2D CNN to learn the global independence of different embeddings. The SASV EER of 2D CNN with SE attention even achieves 0.998% on evaluation set as well as SPF EER and SV EER achieve the best performance compared with other models in Table 1.

Table 1: *EER% results (trained with train 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_SEA	0.876	0.110	1.699	**1.117**	0.519	1.638
F	1D_CNN_GA	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_SEA	0.846	0.135	2.167	**0.998**	**0.497**	1.582

Table 2: *EER% results (trained with train 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_SEA	0.078	8.97e-5	0.134	**0.812**	0.570	**0.981**
F_Aug	1D_CNN_GA	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_SEA	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 & H_Aug & I_Aug	Bosaris (Linear Regression)	-	-	-	0.838	0.838	1.136
D_Aug & G_Aug & H_Aug & I_Aug	Average Score	0.067	0.067	0.135	**0.559**	**0.354**	**0.857**

To further improve the experimental results of the proposed systems, we add the ASVspoof 2019 development set into training set to train the above models. The results are illustrated in Table 2. The SASV EER of Extend (512) DNN has 1.821% absolute reductions compared with only the train set of ASVspoof 2019 training. After augmented with the ASVspoof 2019 development set, 1D CNN model, 1D CNN with SE attention model, 1D CNN with global attention model, 2D CNN model, 2D CNN with SE attention model all achieve that SASV EERs of evaluation set are less than 0.991%. Meanwhile, both the SPF EER and SV EER have further reductions. The state of the art model is the 2D CNN with SE attention whose SASV EER, SPF EER and SV EER are 0.734%, 0.416% and 1.104%. The results demonstrate the proposed global attention and circulant matrix transformation have significant contributions for the backend ensemble of speaker verification and countermeasure systems.

Finally, we fuse the above models trained with ASVspoof 2019 train set and development set. We have tried two fusion methods which are linear regression in Bosaris toolkit [30] and averaging scores respectively. The fusion results are showed in Table 3. After averaging the scores of the 1D CNN model, 1D CNN with global attention model, 2D CNN model and 2D CNN with SE attention model and 2D CNN with VSE attention, the SASV EER, SPF EER and SV EER reach to 0.559%, 0.354%, 0.857%. The linear regression is worse than averaging scores because ASVspoof 2019 development set is used to train the models, which leads to the model overfitting to the development set.

6. Conclusions

In this paper, we explore the backend ensemble of ASV and CM systems, which are trained with pretrained speaker embeddings and CM embeddings to make a decision of speaker verification. We adopt different embedding mixture methods, neural network frameworks and enhanced attention mechanisms to mine speaker and detecting characteristics in different embeddings to make our ensemble models with spoofing aware verification capacity. In particular, our proposed global attention in 1D CNN and circulant matrix transformation make significant improvements on the SASV EER of the evaluation set.

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] B. Desplanques, J. Thienpondt, and K. Demuynck, "PECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech*, 2020.

[3] J. P. Giraldo, S. Lauwereins, K. Badami, H. Van Hamme, and M. Verhelst, "18μw soc for near-microphone keyword spotting and speaker verification," in *2019 Symposium on VLSI Circuits*. IEEE, 2019, pp. C52–C53.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *Proc. Interspeech*, 2019.

[9] 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," *arXiv preprint arXiv:2109.00537*, 2021.

[10] 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," *arXiv preprint arXiv:2109.00535*, 2021.

[11] 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," *arXiv preprint arXiv:2202.08433*, 2022.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[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] 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.

[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] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification."

[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] R. M. Gray, "Toeplitz and circulant matrices: A review," 2006.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[27] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*. IEEE, 2017, pp. 5220–5224.

[28] 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.

[29] L. Zhang, Q. Wang, and L. Xie, "Duality temporal-channel-frequency attention enhanced speaker representation learning," *ASRU*, 2021.

[30] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.