The DKU-OPPO System for the 2022 Spoofing-Aware Speaker Verification Challenge

Xingming Wang[1,2]*, Xiaoyi Qin*[1,2]*, Yikang Wang*[2]*, Yunfei Xu*[3]*, Ming Li*[1,2]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Data Science Research Center, Duke Kunshan University, Kunshan, China
[3]Guangdong OPPO Mobile Telecommunications Corp., Ltd., Guangzhou, China

ming.li369@dukekunshan.edu.cn

Abstract

This paper describes our DKU-OPPO system for the 2022 Spoofing-Aware Speaker Verification (SASV) Challenge. First, we split the joint task into speaker verification (SV) and spoofing countermeasure (CM) these two tasks which are optimized separately. For ASV systems, four state-of-the-art methods are employed. For CM systems, we propose two methods on top of the challenge baseline to further improve the performance, namely Embedding Random Sampling Augmentation (ERSA) and One-Class Confusion Loss(OCCL). Second, we also explore whether SV embedding could help improve CM system performance. We observe a dramatic performance degradation of existing CM systems on the domain-mismatched Voxceleb2 dataset. Third, we compare different fusion strategies, including parallel score fusion and sequential cascaded systems. Our submitted cascaded system obtains a **0.21%** SASV-EER on the challenge official evaluation set.

Index Terms: Anti-spoofing, Speaker verification, Spoofing Countermeasure, Spoofing-Aware Speaker Verification

1. Introduction

Although audio spoofing countermeasure (CM) [1] is highly related to Automatic Speaker Verification (ASV)[2], most of the research on these two tasks has been carried out independently in recent years. This may lead to CM systems not being well suited to some ASV scenarios due to overfitting or domain mismatch [3]. To address this gap, the organizers of the ASVspoof Challenge [4, 5] proposed the tandem detection cost function (t-DCF) metric [6], which is highly correlated to both the ASV system and the CM system, to replace the Equal Error Rate (EER) metric, which relied only on the CM system itself. However, the ASVspoof Challenge still focuses on designing and optimizing a stand-alone CM system to calculate the min t-DCF metric in combination with a given official black-box ASV system. This prevents participants from improving the overall performance by enhancing the ASV system or leveraging joint optimization. Therefore, the first spoofing-aware speaker verification (SASV) challenge [7], which aims to promote the development of integrated systems that can perform both ASV and CM tasks, has been organized this year. The goal of this challenge is to build a hybrid system that can detect both zero-effort impostor access attempts and spoofing attacks simultaneously.

The 2022 SASV challenge focuses on logical access spoofing attacks (LA), such as text-to-speech (TTS) and voice conversion (VC), rather than physical access spoofing attacks (PA), such as record and play-back. Due to the lack of large scale dataset with both speaker and spoofing labels available, few studies involving joint ASV and CM optimization have been conducted in the past. Two categories of jointly optimized solutions have been summarized in the evaluation plan [7]. The first is ensemble systems based on a fusion of separate ASV and CM systems. Gomez et al. [8] use an embedding concatenation strategy to construct an ensemble classification system. Another approach is to build a single integrated system. Li et al. [9] propose a single model using multi-task learning with contrastive loss function.

We investigated and implemented different dual-system score combination methods, including score-fusion and cascaded system. Since the ensemble kind of solutions highly relies on the performance of the pre-trained subsystems, we investigate two innovative schemes to improve the CM system performance. The one-class confusion loss aims to reduce the intra-class Euclidean distance of bonafide audio embeddings. The random embedding sampling augmentation mechanism is also proposed for improving the model's generalization to unseen attacks. Moreover, we found that the performance of CM systems trained on ASVspoof 2019 LA data degraded substantially on Voxceleb2, making the CM system more challenging to utilize the speaker embeddings trained by Voxceleb2 data. Among our explorations, the cascaded system still achieves the lowest error rate without considering the computational cost and real-time latency.

The rest of this paper is organized as follows. In section 2, our submitted system for the SASV challenge is represented, which mainly focuses on the CM system structure and score combination strategies. Implementation details in terms of the dataset usage and model hyperparameters are provided in Section 3. Section 4 describes and discusses the results based on our experiments. Conclusions are provided in Section 5.

2. System Description

2.1. ASV subsystem

Four speaker verification subsystems with different network structures have been adopted in our experiments, including ResNet [10], SE-ResNet [11], SimAM-ResNet [12] and ECAPA-TDNN [13]. The global statistic pooling (GSP) is used for the ResNet subsystem while the attentive statistics pooling (ASP) [14] is used for the other three apporaches. The ArcFace[15] which could increase intra-speaker distances while ensuring inter-speaker compactness is utilized as the classifier.

2.2. CM subsystem

This subsection describes the basic network structure of our CM subsystems and the two proposed methods, namely one-class

Corresponding Author: Ming Li.

confusion loss and embedding random sampling augmentation.

2.2.1. Basic network architecture

We choose AASIST [16] which is the challenge CM baseline as our backbone network. It contains a RawNet2 [17] based encoder and an attention network based graph module. AASIST utilizes raw waveforms as the input to learn meaningful high-dimensional spectro-tempora feature maps and then extract graph nodes of feature maps in temporal and frequency domains respectively [16]. With a stack node that learns information from all nodes, the final CM embedding is attained by concatenating various nodes' mean and maximum values. Moreover, as mentioned in [18], Tak et al. provide an improved architecture in which the max pooling layer of the encoded feature maps is replaced by a 2D self-attentive pooling [19], named as AASIST-SAP.

2.2.2. One-Class Confusion Loss function

Although the basic models, AASIST and AASIST-SAP, obtain great results in the development and evaluation set, there is still a large performance gap since there are unseen attack algorithms in the evaluating set. Therefore, it is necessary to enhance the generalization of our system. The space of bonafide audios is relatively stable while the domain of the attack algorithms are very diverse and unpredictable. Inspired by one-class learning [20, 21], we proposed a One-Class Confusion Loss (OCCL), which is similar to the pairwise confusion loss defined in [22].

The binary cross-entropy loss can be defined as follows:

$$\mathcal{L}_{ce} = \sum_{i} -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

where $y_i \in \{0, 1\}$ is the class label and p_i is the probability output of the classifier. The anti-spoofing CM model is trained using a combined objective with the cross-entropy loss and the proposed one-class confusion loss, which is defined as:

$$\mathcal{L}_{occ} = \sum_{i} \sum_{j \neq i} ||e_i - e_j||^2$$

where e_i denotes the embedding vector extracted from the bonafide audios. The purpose of this loss function is to make the Euclidean distance of all bonafide samples more compact in the embedding space. The one-class confusion loss is only applied on bonafide audios during the training process. Therefore, the final combination loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{occ}$$

where lambda is a constant hyperparameter.

2.2.3. Embedding Random Sampling Augmentation

Considering that the evaluation set contains many unseen logical attacks [23], we propose a fine-tuning Embedding Random Sample Augmentation (ERSA) strategy that aims to improve the robustness of the model for unknown scenarios inspired by [24]. The key idea is to generate random embedding samples from the Gaussian distributions with boundary spoof embedding centers as the mean, and labeled as spoofed speech. Firstly, we initialize the embedding centers of bonafide audio and each type of spoofing audio in the development set separately based on the pre-trained model. The boundary embedding center for each type of spoofing audio are defined as the average of the bonafide embedding center and spoofing embedding centers. During fine-tuning, the boundary embedding centers are dynamically updated based on the embeddings of current iteration. Then we randomly generated samples from $N(\hat{\mu}, \Sigma)$ where N is a Gaussian distribution, $\hat{\mu}$ is the boundary spoof embedding center and Σ is the covariance matrix of spoofing embeddings calculated in advance. After each 5 epoch training, the mean and covariance matrix of embedding centers are updated. A detailed algorithm description can be found in [25].

2.2.4. Integrating Speaker Verification Embeddings

In addition, we also explore the possibility of integrating SV embeddings into the CM model. In this case, the final CM embedding is obtained by concatenating the original CM embedding and the SV embedding together followed by two linear layers.

2.3. System Combination

2.3.1. Score Fusion System

The baseline 1 provided by the challenge organizers just generates the final SASV score by a simple score summation. However, the score distribution of the SV system and the CM system are quite different. Thus, we explore two different strategies to combine the scores, namely the normalized score multiplication approach the same way as in [25, 26] and the score calibration and fusion approach based on the Bosaris toolkit [27].

2.3.2. Cascaded System

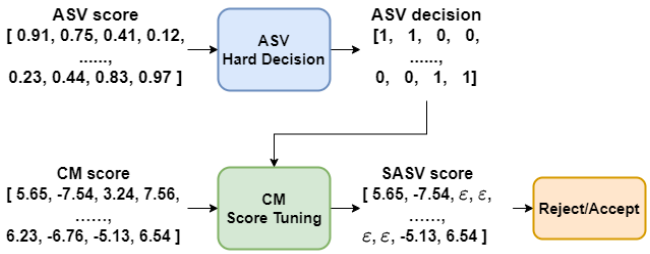

Figure 1: *The illustration of the ASV followed by CM cascaded system. ε represents the minimum CM score in the development set. The CM followed by ASV cascaded system is built in the same way, but switching the SV and CM systems in the pipeline.*

As is shown in Figure 1, the cascaded system consists of two tandem modules: a) the first module generates a hard decision based on a threshold, the threshold is determined by the Equal Error Rate (EER) on the development set; b) the second module directly outputs the raw scores if the decision of module 1 is positive; c) the second module generates the fixed minimum score on the development set when the decision of module 1 is negative. Therefore, once the test audio is tagged as negative by the first module, the score of the second module is not useful.

For the cascaded system, two designs are considered in this work: the first one is ASV module followed by CM module, named as Cascade-ASV-CM; and the other one is CM module followed by ASV module, named as Cascade-CM-ASV. The thresholds of the first system were determined by the EER criteria on the development set.

3. Experimental setup

3.1. Data Usage and Evaluation Metrics

All datasets we used for training and validation are the training and development set of the ASVspoof2019 [23] LA database and the VoxCeleb 2 [28] as requested by the organizers. The ASVspoof2019 LA database consists of bonafide and spoofing audios. Although the database contains both the speaker and spoofing labels, in general it was only used for anti-spoofing countermeasure due to the low number of speakers. The VoxCeleb 2 database contains 1128246 audios from 6112 speakers and has been widely used for ASV training. However, it is difficult to directly train a multi-speaker TTS or VC system just using Voxceleb2. The official SASV evaluation trial consists of audios from the ASVspoof2019 LA evaluation partition, with unseen logical access spoofing attacks compared with audios in the train and development partitions.

The SASV-EER [7], which represents the EER between target and both nontarget and spoof samples, is the primary metric. SPF-EER and SV-EER are adopted as secondary metrics [7].

3.2. Domain Mismatch between ASVspoof2019 LA and VoxCeleb 2

Although the AASIST-based CM system has excellent performance on the ASVspoof2019 evaluation set, it performs poorly on VoxCeleb 2 based on our experiments. Most audios in VoxCeleb 2 are classified as spoofing ones. We summarize two reasons that may lead to this phenomenon.

1. Most audio in VoxCeleb 2 contain various kinds of noises and have been coded and transmitted through some codecs and channels.
2. The CM model trained based on the ASVspoof2019 LA dataset may learn the priori information of silent segments[3].

The great domain mismatches between ASVspoof2019 LA and VoxCeleb 2 datasets makes it difficult to improve the performance of the CM system using the VoxCeleb 2 dataset. Hence, we keep those audio files in VoxCeleb 2 that are classified as bonafide by our CM system as the **Vox-sub** dataset with approximately 20000 audio files in total.

3.3. Model setup

3.3.1. ASV subsystem

For feature extraction, logarithmical Mel-spectrogram is extracted by applying 80 Mel filters on the spectrogram computed over Hamming windows of 20ms shifted by 10ms. The on-the-fly data augmentation [29] is employed to add additive background noise or convolutional reverberation noise for the time-domain waveform. The MUSAN [30] and RIR Noise [31] datasets are used as noise sources and room impulse response functions, respectively. We apply amplification or speed change (pitch remains untouched) to audio signals to further diversify training samples. Also, we apply speaker augmentation with speed perturbation [32, 33, 34]. We adopt the Reduceonplateau learning rate (LR) scheduler with 0.1 initial LR. The SGD optimizer is adopted to update the model parameters.

3.3.2. CM subsystem

In contrast to the baseline training strategy, our trained AASIST-SAP network receives random length audio between 3-5 seconds as input. The initial learning rate is 0.001 with a Reduceonplateau learning rate scheduler. Adam optimizer is used to update the weights in models. The embedding random sample augmentation is only used during fine-tuning with two generated embeddings per center. Since there are six boundary spoof embedding centers, there will be 12 generated embeddings per batch. The batch size is set as 64 in this phase. And for the one-class confusion loss, λ is set as 1 during training. For fusing SV embedding into the CM model, considering the great mismatch mentioned in section 3.2, we adopt SV embeddings generated by the Resnet-GSP model trained by Voxceleb2 and Voxsub as **SV-embd-V1** and **SV-embd-V2**, respectively.

4. Results and discussion

4.1. Results of the ASV Subsystem

Table 1 reports the results of different speaker verification models. Our used models achieve state-of-the-art results on the VoxCeleb1 original test set. The ResNet with GSP achieves the best single model performance on the SASV dataset which might be because the generalization of ResNet with statistic pooling is better on this ASVspoof dataset.

Table 1: *The performances of different speaker verification subsystems on the VoxCeleb1 original test set and the SASV challenge dataset.*

Model	**Vox-O EER[%]**	**SV-EER[%]**	
		Dev	Eval
ECAPA (Baseline)	-	1.86	1.64
ResNet GSP	0.851	**0.135**	**0.192**
SE-ResNet34 ASP	0.776	0.404	0.410
SimAM-ResNet34 ASP	0.643	0.404	0.252
ECAPA-TDNN	0.734	0.225	0.228

Table 2: *Comparison of different single CM systems based on SPF-EER used in SASV challenge. The 19LA denotes using both train and dev sets of ASVspoof2019 for training. The Vox-sub represents sub-bonafide audios selected from VoxCeleb 2 as mentioned in Sec 3.2. The SV-embd-V1 and V2 are defined in Sec 3.3.2.*

Model	**Data**	**SPF-EER[%]**	
		Dev	Eval
AASIST(Baseline)	19LA train	0.07	0.67
AASIST	19LA	0.067	0.668
AASIST-SAP	19LA	0.067	0.570
AASIST-SAP+$ERSA$	19LA	0.067	0.510
AASIST-SAP+SV-embd-V1	19LA	0.058	4.320
AASIST-SAP+SV-embd-V2	19LA	0.067	1.229
AASIST-SAP	19LA+Vox-sub	0.049	1.564
AASIST-SAP+$OCCL$	19LA+Vox-sub	**0.000**	**0.360**

4.2. Results of the CM Subsystem

Table 2 shows the anti-spoofing performance of different CM subsystems. It can be seen from the table that the AASIST based model achieves a great performance improvement by replacing the max pooling layer with SAP. In addition, the model

Table 3: *Performance of different systems evaluated in the SASV Challenge. Due to a large number of combinations, only selected combinations are listed. The σ denotes $sigmoid$ normalization and $\times$ denotes multiplication.*

ID	Model	Fusion	SV-EER[%]		SPF-EER[%]		SASV-EER[%]	
			Dev	Eval	Dev	Eval	Dev	Eval
CM System								
1	AASIST(Baseline)	-	46.01	49.24	0.07	0.67	15.86	24.38
2	AASIST-SAP+$ERSA$	-	47.304	47.188	0.067	0.510	15.963	24.655
3	AASIST-SAP+$OCCL$	-	50.644	55.161	**0.000**	**0.360**	16.328	26.872
ASV System								
4	ECAPA-TDNN (Baseline)	-	1.86	1.64	20.28	30.75	17.31	23.84
5	ResNet34 GSP	-	**0.135**	**0.192**	14.084	23.069	11.616	17.449
6	SE-ResNet34 ASP	-	0.404	0.410	11.540	22.402	9.745	16.888
7	ECAPA-TDNN	-	0.225	0.228	14.420	21.899	12.354	16.795
8	SimAM-ResNet34 ASP	-	0.404	0.252	12.011	22.500	10.512	16.994
	Baseline 1 (official)	Sum	32.89	35.33	0.07	0.67	13.06	19.31
	Baseline 2 (official)	Back-end ensemble	7.94	9.29	0.07	0.80	3.10	5.23
Score-fuse	ID 5+6+7 & ID 1+2+3	Sum	19.694	23.706	0.000	0.186	8.630	13.892
	ID 5+6+7 & ID 1+2+3	σ and $\times$	0.202	0.317	0.000	0.186	**0.103**	**0.279**
	ID 5+6+7 & ID 1+2+3	Bosaris	0.134	0.298	0.009	0.577	0.067	0.487
Cascade	ID 5 & ID 1	Cascade-ASV-CM	0.134	0.204	0.202	0.477	0.202	0.428
		Cascade-CM-ASV	0.202	0.335	0.067	0.684	0.149	0.503
	ID 5 & ID 3	Cascade-ASV-CM	0.135	0.205	0.135	0.410	0.135	0.391
		Cascade-CM-ASV	0.135	0.410	0.000	0.298	0.128	0.410
	ID 8 & ID 2	Cascade-ASV-CM	0.404	0.390	0.404	0.260	0.404	**0.288**
		Cascade-CM-ASV	0.451	1.359	0.067	1.252	0.202	1.322
	ID 5+6+7 & ID 1+2+3	Cascade-ASV-CM **(submitted)**	0.202	0.462	0.202	0.186	0.202	**0.209**
		Cascade-CM-ASV	0.173	0.242	0.000	0.230	**0.096**	0.242

achieves a further generalizability improvement in the evaluation set by fine-tuning with ERSA strategy.

It is worth mentioning that although we extracted the Voxsub dataset, simply adding these bonafide samples to the CM training set does not improve the CM performance. However, after integrating the proposed OCCL loss function, the overall CM performance is further enhanced. This improvement may be attributed to the fact that this loss function makes the Euclidean distances between embeddings of bonafide audios in VoxCeleb 2 and ASVSpoof2019 LA closer, and thus the bonafide embedding space is more compact.

Unfortunately, simply fusing SV embedding into the CM system seems not useful. This may still be due to the domain mismatch mentioned in section 3.2, as the CM performance with SV-embd-V2 has increased considerably compared with the one with SV-embd-V1.

4.3. Results on the Combined System

The results of score ensemble experiments are summarized in Table 3. More detailed results can be found in [25].

As can be seen from the score fusion section of the table, the simple summation method performs poorly due to the differences among the score distributions of different subsystems. This problem can be effectively mitigated by normalizing the scores through the $sigmoid$ function and multiplying them together [26]. The optimal result in this part is also obtained by this method.

The cascaded systems section of the table shows results of different cascading combinations. We have noted that the AASIST CM system in the baseline is highly complementary to the systems trained by ourselves. Furthermore, while the Cascade-CM-ASV approach performed better on the development set, the Cascade-ASV-CM approach generally performed better on the evaluation set, possibly because the development set has appeared in the training data of CM systems. In other words, for unknown scenarios, the more generalized system with lower EER is more suitable to be the first module with hard decisions. Finally, we submit the results of the Cascade-ASV-CM method.

5. Conclusion

In this paper, we describe our submitted system for the 2022 SASV challenge. We mainly focus on the CM subsystem and propose an embedding random sampling fine-tuning strategy to improve performance. Besides, by considering the great domain mismatch between datasets, we propose the one-class confusion loss, which improves the CM subsystem's performance even further. The final cascaded system submitted achieved 0.21% EER on the SASV challenge evaluation set. In the future, we will try to collect a large scale database with both speaker and spoofing labels available. So we can explore more advanced joint learning approaches with sufficient data.

6. Acknowledgements

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

7. References

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

[2] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[3] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?" in *Proc. ASVspoof2021 workshop*, 2021, pp. 55–60.

[4] 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.

[5] 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. ASVspoof2021 workshop*, 2021, pp. 47–54.

[6] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification ," in *Proc. Speaker Odyssey*, 2018, pp. 312–319.

[7] 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.

[8] 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.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. CVPR*, 2018.

[12] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*, 2022.

[13] D. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[14] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," *Proc. Interspeech*, 2018.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

[16] 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," in *Proc. ICASSP*, 2022.

[17] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," in *Proc. ICASSP*, 2021, pp. 6369–6373.

[18] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[19] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[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] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation," *Proc. ASVspoof2021 workshop*, pp. 16–21, 2021.

[22] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proc. ECCV*, 2018, pp. 70–86.

[23] 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.

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

[25] X. Wang, X. Qin, Y. Wang, Y. Xu, and M. Li, "The DKU-OPPO System Description for the First SASV Challenge: Unknown Attack Samples Detection Using Embedding Augmentation and One-class Confusion Loss," https://sasv-challenge.github.io/pdfs/2022_descriptions/DKU-OPPO.pdf.

[26] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," *arXiv preprint arXiv:2202.05253*, 2022.

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

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[29] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484*.

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

[32] H. Yamamoto, L. K.A., K. Okabe, and T. Koshinaka, "Speaker Augmentation and Bandwidth Extension for Deep Speaker Embedding," in *Proc. Interspeech*, 2019, pp. 406–410.

[33] W. Wang, D. Cai, X. Qin, and M. Li, "The DKU-DukeECE Systems for VoxCeleb Speaker Recognition Challenge 2020," *arXiv.2010.12731*.

[34] X. Qin, C. Wang, Y. Ma, M. Liu, S. Zhang, and M. Li, "Our Learned Lessons from Cross-Lingual Speaker Verification: The CRMI-DKU System Description for the Short-Duration Speaker Verification Challenge 2021," in *Proc. Interspeech*, 2021, pp. 2317–2321.