Interspeech 2022
18-22 September 2022, Incheon, Korea

Spoofing-Aware Speaker Verification by Multi-Level Fusion

Haibin Wu[1,2], *Lingwei Meng*[3], *Jiawen Kang*[3], *Jinchao Li*[3], *Xu Li*[3], *Xixin Wu*[3], *Hung-yi Lee*[1], *Helen Meng*[2,3]

[1] Graduate Institute of Communication Engineering, National Taiwan University
[2] Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong
[3] Human-Computer Communications Laboratory, The Chinese University of Hong Kong

Abstract

Recently, many novel techniques have been introduced to deal with spoofing attacks, and achieve promising countermeasure (CM) performances. However, these works only take the stand-alone CM models into account. Nowadays, a spoofing aware speaker verification (SASV) challenge which aims to facilitate the research of integrated CM and ASV models, arguing that jointly optimizing CM and ASV models will lead to better performance, is taking place. In this paper, we propose a novel multi-model and multi-level fusion strategy to tackle the SASV task. Compared with purely scoring fusion and embedding fusion methods, this framework first utilizes embeddings from CM models, propagating CM embeddings into a CM block to obtain a CM score. In the second-level fusion, the CM score and ASV scores directly from ASV systems will be concatenated into a prediction block for the final decision. As a result, the best single fusion system has achieved the SASV-EER of 0.97% on the evaluation set. Then by ensembling the top-5 fusion systems, the final SASV-EER reached 0.89%.

Index Terms: anti-spoofing, speaker verification, spoofing-aware speaker verification

1. Introduction

Spoofing attacks and countermeasures for automatic speaker verification (ASV) have aroused keen interests in academia and the industry. While ASV systems aim to verify the identity of target speakers, spoofing attacks attempt to manipulate the verification results using various technologies, leading to dramatic performance degradation [1–4]. In order to ensure the robustness and security of ASV systems, CM is a necessary technique to detect and defend spoofing attacks.

The vulnerability of ASV systems is vulnerable to speech synthesis and voice conversion (VC) attacks [1–3], and recently emerged adversarial attacks [5]. Various methods have been proposed to tackle adversarial attacks [6–11]. Currently, various techniques have been proposed to perform effective attacks, including audio replay [12], adversarial noise [5, 13], more advanced text-to-speech (TTS) and VC models [14–16]. In addition, many works have been done to investigate state-of-the-art CM strategies. The current solutions leverage end-to-end deep neural networks (DNNs) [17, 18], trying to distinguish artifacts and unnatural cues of spoofing speech from bona fide speech. And thanks to a series of challenges and datasets [1–3], many novel techniques were introduced to achieve promising CM performances [19–31].

However, previous works only take the stand-alone CM models into account. Recently, a spoofing aware speaker verification (SASV) challenge [32] was proposed as a special session in ISCA INTERSPEECH 2022. This challenge aims to facilitate the research of integrated CM and ASV models, arguing that jointly optimizing CM and ASV models will lead to better performance. To measure the performance of integrated models, a SASV-EER was proposed in this challenge as a primary metric, which is a variant of classic equal error rate (EER). Under this metric, the test utterances in trials belong to one of three types: impostors, target speakers, and spoofing attacks. In further, the SASV-EER can be subsetted into SV-EER (impostors vs. targets) and SPF-EER (targets vs. spoof). The former is for evaluating speaker verification performance, and the latter is for evaluating anti-spoofing performance. In this way, this metric expects the model can accept target speakers and reject any alternatives, including the impostors and spoofing attacks.

This paper described our submitted system for the SASV Challenge 2022. In order to take advantage of existing well-designed models in CM and ASV areas, we proposed a novel multi-model and multi-scale fusion framework. Compared with purely scoring fusion and embedding fusion methods, this framework utilizes embeddings from CM models in the first-level fusion, propagating CM embeddings into a CM block to obtain a CM score. In the second-level fusion, the CM score and ASV scores directly from ASV systems will be concatenated into a prediction block for the final decision. In contrast to our previous work [33] that only simply concatenates the embeddings from different CM models, we considered the potentials of pooling strategies in terms of feature aggregation, and investigated various pooling methods [34–37] when fusing embeddings across different CM models. Based on the proposed fusion framework, we presented the fusion strategies of a series of state-of-the-art CM and ASV models with different pooling strategies to boost the fusion results. As a result, the best single fusion system has achieved the SASV-EER of 0.97% on the evaluation set. Then by ensembling the top-5 fusion systems, the final SASV-EER reached 0.89% on the evaluation set, while this number in the best baseline system from the SASV challenge is 6.37%.

2. Background

2.1. SASV strategies

Given the enrollment utterance x_e and the testing utterance x_t, spoofing-aware speaker verification (SASV) systems aim at telling $y_t = 1$ if x_t comes from the same speaker as x_e, or $y_t = 0$ if x_t comes from another speaker or x_t is a spoofing attack. There are two typical strategies for constructing a SASV system: multi-task learning strategy and fusion-based strategy.

The multi-task learning strategy trains the models jointly with both speaker verification and anti-spoofing objectives, which is intuitive to be adopted. The two objectives share the same backbone and thus the features and embeddings, while each objective has its own predicting head and loss function. It is worth noting that speaker verification and anti-spoofing ob-

Copyright © 2022 ISCA

10.21437/Interspeech.2022-920

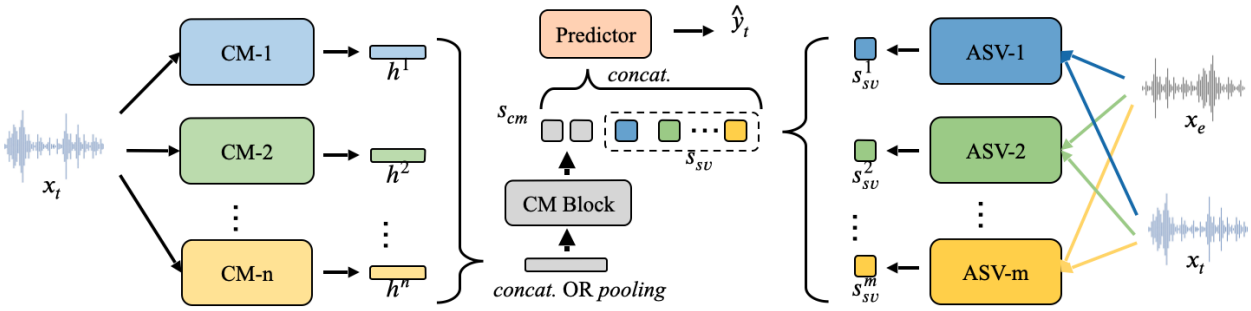


Figure 1: *The proposed multi-model & multi-level fusion framework.*

jectives are contradictory in some respects. The former drives the model to erase device and environment information to more robustly identify speakers; in contrast, the latter prompts the model to capture device and environment traces, then tells forged spoofing from authentic utterances [24].

Alternatively, the fusion-based strategy has the potential to reach better SASV performance leveraging state-of-the-art CM and ASV models trained on large-scale datasets. Considering this superiority, we propose our solutions for the SASV challenge based on a novel multi-model and multi-level fusion strategy.

2.2. Baseline systems

The challenge organizer provides two baseline systems. Each system is based upon ECAPA-TDNN model [38] as the ASV subsystem and AASIST model [21] as the CM subsystem. The key difference between the two systems is how they fuse ASV and CM subsystem - Baseline1 adopts the score-level fusion, while Baseline2 adopts the embedding-level fusion.

3. Method

Although achieving acceptable performances, baseline systems' fusion strategies are relatively simple and naive. Baseline1's score-level fusion does not guarantee that the ASV score and the CM score belong to a unified space and an identical magnitude. Baseline2 crudely concatenates three embeddings and throws the product into a DNN, lacking fine-grained fusion. These inadequacies motivate us to explore further possibilities of fusion strategies in the SASV task.

In this paper, we proposed a multi-model and multi-level fusion strategy. In terms of the width, it employs multiple pre-trained ASV and CM models as plug-and-play components, where users can expand or shrink the scale of the model according to their needs. In terms of the depth of this strategy, it fuses CM embeddings to calculate a score in the first-level fusion, which is then integrated with the ASV models' outputs, and yields the final prediction in the second-level fusion.

3.1. Overall structure

The overall framework is shown in Figure 1, where x_e and x_t are the input enrollment and testing utterances, respectively; *ASV-1*, *ASV-2*, ..., *ASV-m* denote m pre-trained ASV models; *CM-1*, *CM-2*, ..., *CM-n* denote n pre-trained CM models. Given a trial $\{x_e, x_t\}$, a series of cosine scores $\{s^1_{sv}, s^2_{sv}, \ldots, s^m_{sv}\}$ are derived from m ASV models. Given the testing utterance x_t, a series of CM embeddings $\{h^1, h^1, \ldots, h^n\}$ are extracted through n CM models. Next comes the first level of fusion, where the n embeddings are integrated into a h_{cm} by concatenation or a pooling method. Our previous work [33] investigated the capacity of concatenation for SASV. In this work, we further extend the potential of concatenation by making the CM block in Figure 1 deeper, and we also adopt pooling methods to further improve the fusion performance. Further, h_{cm} goes through a CM Block to better digest fused embeddings and then a 2-dimension countermeasure score s_{cm} is predicted. With s_{cm} and $\{s^1_{sv}, s^2_{sv}, \ldots, s^m_{sv}\}$ well prepared, they are concatenated and fed into the Predictor to yield the final prediction $\hat{y}_t$, where the second-level fusion is performed.

3.2. Strategies in the first-level fusion

For the first-level fusion, we attempt concatenation [33] or one of the four kinds of pooling methods to synthesize h_{cm} separately, and have conducted extensive experiments accordingly. Suppose $\boldsymbol{H} = \{h^1, h^1, \ldots, h^n\}$, and the length of all the CM embeddings are projected into a same length d_h by feed forward layers before performing concatenation or pooling. The four candidate pooling methods are Temporal Average Pooling (TAP) [34], Temporal Statistics Pooling (TSP) [35], Self-attentive Pooling (SAP) [36], and Attentive Statistics Pooling (ASP) [37].

TAP is to calculate the mean value along the channels to obtain the h_{cm}.

TSP calculates channel-wise mean and standard deviation, then concatenates the mean vector and standard deviation vector together as h_{cm}.

In SAP, the self-attention mechanism takes $\boldsymbol{H}$ as input and outputs an annotation matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = softmax(tanh(\boldsymbol{H}^T \boldsymbol{W}_1)\boldsymbol{W}_2) \tag{1}$$

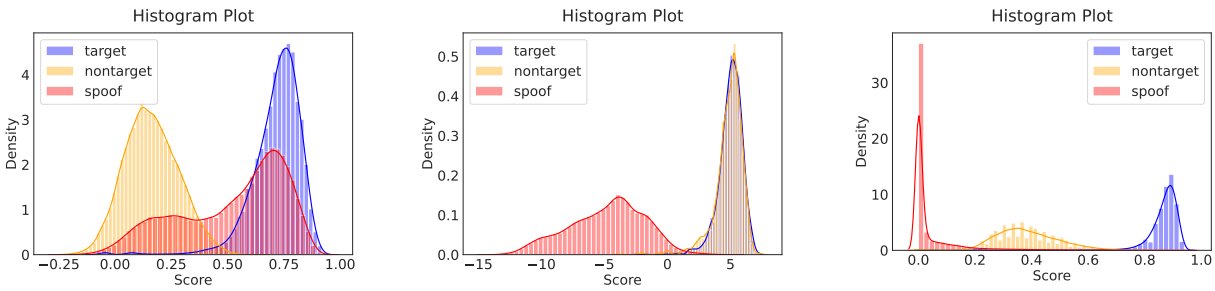
where $\boldsymbol{W}_1$ is a matrix of size $d_h \times d_a$; $\boldsymbol{W}_2$ is a matrix of size $d_a \times d_r$, and d_r is a hyper-parameter that represents the number of attention heads; The $softmax()$ is performed column-wise. We set $d_r = 1$, therefore $\boldsymbol{A}$ degenerates into an annotation vector. Weighted by $\boldsymbol{A}$, h_{cm} is calculated as the weighted mean:

$$h_{cm} = \tilde{\mu} = \boldsymbol{H}\boldsymbol{A} \tag{2}$$

For ASP, not only it calculate a attention-weighted mean as SAP do, but also it calculate a attention-weighted standard deviation:

$$\tilde{\sigma} = \sqrt{\sum_{i=1}^{n} \alpha^i h^i \odot h^i - \tilde{\mu} \odot \tilde{\mu}} \tag{3}$$

where α^idenotes the i^{th} element of the annotation vector $\boldsymbol{A}$, $\odot$ represents the Hadamard product. By concatenating $\tilde{\mu}$ and $\tilde{\sigma}$, h_{cm} is derived.



(a) ECAPA-TDNN (b) AASIST (c) Proposed top-5 ensemble system

Figure 2: *The histogram plots of the output scores predicted by ASV, CM, and the proposed top-5 ensemble system. Other proposed system variants also have similar histogram shapes as (c).*

After one of the above pooling methods or concatenation, derived h_{cm} goes through CM Block, which is a multi-layer perceptron, and generates a two-dimension score reflecting the possibilities the testing utterances is the target or not.

3.3. Loss function

The ASV and CM models parameters are well pre-trained and thus frozen. The learnable modules include CM Block and the Predictor, which are multi-layer perceptrons. To prompt the model to learn to distinguish the target trials from other trials, we use the cross-entropy loss on s_{cm} output by CM Block and $\hat{y}_t$ output by the Predictor respectively.

4. Experiments

4.1. Experimental setup

In the SASV Challenge 2022 [32], participants are restricted to utilise ASVspoof 2019 [39] and VoxCeleb2 [40] datasets for training the anti-spoofing and ASV model, respectively.

Three EERs, namely SV-EER, SPF-EER and SASV-EER are measured as the evaluation metrics, and SASV-EER is the main metric in the Challenge.

For the ASV models, we use Resnet34 [41], ECAPA-TDNN [38] and MFA-Conformer [42]. For the countermeasure models, we use AASIST [21], AASIST-L, and RawGAT-ST [28], where AASIST-L is a light version of AASIST. The fusion model in Figure 1 is trained by Adam optimizer with an initial learning rate of 0.0001. We set the batch size as 32, and the epoch number as 40.

4.2. Experimental results and analysis

As the requirements by SASV Challenge 2022, we evaluated systems on ASVspoof 2019 development and evaluation sets and reported SA-EER, SPF-EER and SASV EER, shown in Table 1. A and B denote systems using pure ASV or CM. We omit other systems using only single ASV or CM due to space limitation, yet they are with the same trend. C1-C2 denote two baselines provided by the Challenge organizer. D-H8 are variants based on the proposed fusion strategy. D denote the fusion systems using only one ASV model and one CM model, e.g., 'ECAPA-TDNN + AASIST' denotes the fusion of ECAPA-TDNN as the ASV model, and AASIST as the CM model. Other systems using only one ASV and CM models are with the same trend. E1-E3 denote systems fusing all three ASV models with one CM model. Note that D-E3 ignore the first-level fusion. F1-F3 denote systems fusing one ASV model and all three CM models. G1-H8 represent systems incorporating all three ASV models and all three CM models, but with different first-level fusion strategies and different sizes of CM Block. For example, 'SV-ALL + CM-ALL-CAT-768' denotes its first-level fusion uses concatenation (abbreviated as 'CAT' in the table), and fused h_{cm} is projected to 768 dimensions in CM Block's first layer. I1 is the ensemble system involving the top-5 best evaluation set SASV-EER systems in A-H8. Figure 1 illustrates the histogram plots of three typical systems: (a) ECAPA-TDNN, a SOTA ASV system; (b) AASIST, a SOTA CM system; (c) Proposed top-5 ensemble system.

4.2.1. Single-objective systems

Only using speaker verification models. ECAPA-TDNN, MFA-Conformer, Resnet34 performed well on the speaker verification sub-task and achieved 1.86%, 1.38% and 1.08% SV-EERs on the evaluation set, respectively. However, they perform unacceptably on the anti-spoofing sub-task, yielding 30.75%, 30.22%, 29.76% SPF-EER. Spoiled by the spoofing attacks, it is unpractical to perform SASV tasks using only the ASV models. Take ECAPA-TDNN model as an example, as shown in Figure 2 (a), the pure ASV model can separate target and non-target trials while can hardly distinguish spoofing trials from genuine trials. This phenomenon is predictable because the objective of speaker verification models tends to erase the in-congruent device and environment information to more robustly identify speakers. However, these real-world traces should have helped to defend against spoofing attacks.

Only using anti-spoofing models. In contrast, AASIST, the state-of-the-art CM model, can significantly discriminate spoofing utterances, but randomly guess on the speaker identification sub-task, mainly because of their speaker-unrelated objectives. AASIST, AASIST-L and RawGAT-ST achieve SPF-EERs of 0.67%, 0.84%, 0.96% on anti-spoofing, while yielding SV-EERs close to 50% for speaker verification. This phenomenon can also be observed in Figure 2 (b), where the target and non-target trials' distribution are almost totally overlapped.

4.2.2. Baseline systems

Compared to the above single-objective models, two baseline systems reveal superiority in the SASV challenge. Among them, Baseline2 achieves a SASV-EER of 8.75% on the evaluation set, which is better than Baseline1's 19.15%. We argue that the reason is that the simple score-level fusion used in Baseline1 does not guarantee that ASV and CM scores belong to a unified

Table 1: *Performance of all systems on the ASVspoof 2019 development and evaluation sets.*

	System	SV-EER		SPF-EER		SASV-EER	
		Dev	Eval	Dev	Eval	Dev	Eval
(A)	ECAPA-TDNN	1.64	1.86	20.28	30.75	17.37	23.84
(B)	AASIST	46.01	49.24	0.07	0.67	15.86	24.38
(C1)	Baseline1 [32]	32.88	35.32	0.06	0.67	13.07	19.31
(C2)	Baseline2 [32]	12.87	11.48	0.13	0.78	4.85	6.37
(D)	MFA-Conformer + AASIST	1.48	1.47	0.20	1.08	0.88	1.35
(E1)	SV-ALL + AASIST	1.42	1.30	0.27	1.61	0.81	1.41
(E2)	SV-ALL + AASIST-L	1.42	1.33	0.47	3.99	0.88	2.95
(E3)	SV-ALL + RawGAT-ST	1.82	1.64	0.40	0.82	1.28	1.39
(F1)	MFA-Conformer + CM-ALL-CAT-256	1.91	1.66	0.20	0.64	1.01	1.30
(F2)	ECAPA-TDNN + CM-ALL-CAT-256	1.39	1.73	0.20	0.74	0.81	1.40
(F3)	Resnet34 + CM-ALL-CAT-256	1.28	1.12	0.26	1.43	0.74	1.32
(G1)	SV-ALL + CM-ALL-CAT-256	1.27	1.20	0.20	1.15	0.81	1.17
(G2)	SV-ALL + CM-ALL-CAT-512	1.35	1.15	0.20	1.12	0.74	1.14
(G3)	SV-ALL + CM-ALL-CAT-768	1.34	1.12	0.20	0.99	0.81	**1.08**
(G4)	SV-ALL + CM-ALL-CAT-1024	1.28	1.21	0.20	0.83	0.74	**1.08**
(G5)	SV-ALL + CM-ALL-CAT-2048	1.35	1.10	0.23	1.41	0.74	1.31
(H1)	SV-ALL + CM-ALL-TAP-768	1.35	0.99	0.2	1.10	0.61	**1.02**
(H2)	SV-ALL + CM-ALL-TSP-768	1.21	1.12	0.20	1.47	0.74	1.31
(H3)	SV-ALL + CM-ALL-SAP-768	1.15	1.04	0.17	0.93	0.54	**0.99**
(H4)	SV-ALL + CM-ALL-ASP-768	1.18	1.37	0.15	1.58	0.67	1.51
(H5)	SV-ALL + CM-ALL-TAP-1024	1.28	1.15	0.13	0.56	0.61	**0.97**
(H6)	SV-ALL + CM-ALL-TSP-1024	1.11	1.12	0.20	1.77	0.61	1.43
(H7)	SV-ALL + CM-ALL-SAP-1024	1.15	1.16	0.20	1.45	0.61	1.28
(H8)	SV-ALL + CM-ALL-ASP-1024	1.51	1.68	0.40	1.01	1.08	1.49
(I1)	Top-5 Ensemble	1.08	1.01	0.20	0.71	0.67	**0.89**

space with a consistent magnitude. As shown in Figure 2, the ASV score ranges from -1 to 1, while the CM score ranges from -20 to 15. Straightforward addition will make the ASV score submerged by the CM score. In comparison, Baseline2 uses a trainable deep neural network to digest ASV and CM embeddings better, which can contribute to distinguishing the target from non-target and spoofing trials. However, the performance of Baseline2 on the speaker verification sub-task is still not satisfactory, which motivates us to propose the multi-model and multi-level fusion strategy.

4.2.3. Proposed fusion systems

The proposed strategy fuses three SOTA ASV models and three SOTA CM models. With different settings, variety of variant fusion systems (D-H8) are elaborately designed, and an additional ensemble system (I1) is constructed by integrating the top-5 best evaluation set SASV-EER systems from A-H8. Table 1 illustrates the performance of all systems, and G1 is the best result of our previous work [33].

From Table 1, all the proposed systems outperform the baseline models with a large margin on the SASV-EER metrics, while retaining universally good performances on SV-EER and SPF-EER. The proposed models perform consistently well on the speaker verification task, and quite a few can reach or even surpass the performance of the SOTA ASV models. G1-H8 show better performance than D-F3 in general, benefiting from the more comprehensive model fusion.

Using concatenation for the first-level fusion, we attempt different CM Block sizes. From G1 to G5, as we increase the size of the CM Block, the SASV-EER of the system decreases. Until adding a layer with an output dimension of 2048 to the bottom of the CM Block, the performance saturates.

Since G3 and G4 have the best SASV-EER, we employ their CM Block settings and further explore the impact of replacing concatenation with different pooling methods. In H1-H8, under both CM block settings, TAP bring more benefits to the system than other methods. Both statistics pooling methods (TSP, ASP) are less effective than TAP, SAP, and concatenation. A possible reason is that we only compute the standard deviation for up to three CM embeddings, which is not stable during training. Rather than representing useful knowledge, the standard deviation seems to be more of a noise. We argue that if more CM models enroll, the system will benefit from statistics pooling methods, and we will leave it as future work.

The best individual system is with SASV-EER as 0.97%. Moreover, the top-5 ensemble system (I1) achieves a SASV-EER as low as 0.89%, which is a 86% relative improvement compared to Baseline2. Figure 2 (c) shows that the distributions of the target, non-target and spoofing trials are well-separated, which verifies the effectiveness of the proposed method.

5. Conclusion

This work proposes a novel multi-model and multi-level fusion strategy to tackle the SASV task. The two-level fusion method can take advantage of the state-of-the-art ASV and CM models. As a result, the best single fusion system achieves the SASV-EER of 0.97%. What's more, by ensembling the top-5 systems, the final SASV-EER reaches 0.89% on the evaluation set, which is 86% relative reduction compared to the best baseline, Baseline2, and 24% relative reduction compared to our previous work [33]. In the future work, we will introduce more CM models to investigate the potential of the proposed method.

6. References

[1] T. Kinnunen, M. Sahidullah *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *ISCA (the International Speech Communication Association)*, 2017.

[2] M. Todisco, X. Wang *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[3] J. Yamagishi, X. Wang *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[4] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[5] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *arXiv preprint arXiv:2004.08849*, 2020.

[6] X. Li, N. Li, J. Zhong, X. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Investigating robustness of adversarial samples detection for automatic speaker verification," *arXiv preprint arXiv:2006.06186*, 2020.

[7] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *ICASSP*. IEEE, 2021, pp. 6718–6722.

[8] H. Wu *et al.*, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE/ACM TASLP*, vol. 30, pp. 202–217, 2021.

[9] H. Zhang, L. Wang, Y. Zhang, M. Liu, K. A. Lee, and J. Wei, "Adversarial separation network for speaker recognition," *Proc. Interspeech 2020*, pp. 951–955, 2020.

[10] H. Wu, Y. Zhang, Z. Wu, D. Wang, and H.-y. Lee, "Voting for the right answer: Adversarial defense for speaker verification," *arXiv preprint arXiv:2106.07868*, 2021.

[11] H. Wu *et al.*, "Spotting adversarial samples for speaker verification by neural vocoders," *arXiv preprint arXiv:2107.00309*, 2021.

[12] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *APSIPA*. IEEE, 2014, pp. 1–5.

[13] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *ICASSP*. IEEE, 2018, pp. 1962–1966.

[14] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," 2010.

[15] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[16] Z. Kons and H. Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems." in *INTERSPEECH*, 2013, pp. 945–949.

[17] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *ASRU*. IEEE, 2019, pp. 1003–1010.

[18] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101096, 2020.

[19] 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.

[20] H. Wu *et al.*, "Partially fake audio detection by self-attention-based fake span discovery," *arXiv preprint arXiv:2202.06684*, 2022.

[21] 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," *arXiv preprint arXiv:2110.01200*, 2021.

[22] X. Li *et al.*, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021*. IEEE, 2021, pp. 6354–6358.

[23] H. Tak, J.-w. Jung, J. Patino, M. Todisco, and N. Evans, "Graph attention networks for anti-spoofing," *arXiv preprint arXiv:2104.03654*, 2021.

[24] H.-j. Shim, J.-w. Jung, J.-h. Kim, and H.-j. Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, vol. 10, no. 18, p. 6292, 2020.

[25] J.-w. Jung, H.-j. Shim, H.-S. Heo, and H.-J. Yu, "A study on the role of subsidiary information in replay attack spoofing detection," *arXiv preprint arXiv:2001.11688*, 2020.

[26] H.-j. Shim, H.-S. Heo, J.-w. Jung, and H.-J. Yu, "Self-supervised pre-training with acoustic configurations for replay spoofing detection," *arXiv preprint arXiv:1910.09778*, 2019.

[27] J.-w. Jung, H.-j. Shim, H.-S. Heo, and H.-J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoof 2019 challenge," *arXiv preprint arXiv:1904.10134*, 2019.

[28] H. Tak *et al.*, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[29] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *Interspeech*, 2018, pp. 641–645.

[30] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," *arXiv preprint arXiv:2005.10393*, 2020.

[31] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection." in *INTERSPEECH*, 2018, pp. 726–730.

[32] 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.

[33] H. Wu, J. Kang, L. Meng, Y. Zhang, X. Wu, Z. Wu, H. yi Lee, and H. Meng, "Tackling spoofing-aware speaker verification with multi-model fusion," *arXiv preprint arXiv:2206.09131*, 2022.

[34] L. You, W. Guo, L. Dai, and J. Du, "Multi-task learning with high-order statistics for x-vector based text-independent speaker verification," *arXiv preprint arXiv:1903.12058*, 2019.

[35] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018*. IEEE, 2018, pp. 5329–5333.

[36] 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.

[37] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018, pp. 2252–2256.

[38] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[39] 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.

[40] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, and Z. Wu, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," *arXiv preprint*, 2022.