applied sciences

MDPI

Article

Integrated Replay Spoofing-Aware Text-Independent Speaker Verification

Hye-jin Shim †, Jee-weon Jung †, Ju-ho Kim and Ha-jin Yu *

School of Computer Science, University of Seoul, Seoul 02504, Korea; shimhz6.6@gmail.com (H.-j.-S.); jeewon.leo.jung@gmail.com (J.-w.J.); wngh1187@naver.com (J.-h.K.)

* Correspondence: hjyu@uos.ac.kr

† These authors contributed equally to this work.

Received: 6 August 2020; Accepted: 7 September 2020; Published: 10 September 2020

Abstract: A number of studies have successfully developed speaker verification or presentation attack detection systems. However, studies integrating the two tasks remain in the preliminary stages. In this paper, we propose two approaches for building an integrated system of speaker verification and presentation attack detection: an end-to-end monolithic approach and a back-end modular approach. The first approach simultaneously trains speaker identification, presentation attack detection, and the integrated system using multi-task learning using a common feature. However, through experiments, we hypothesize that the information required for performing speaker verification and presentation attack detection might differ because speaker verification systems try to remove device-specific information from speaker embeddings, while presentation attack detection systems exploit such information. Therefore, we propose a back-end modular approach using a separate deep neural network (DNN) for speaker verification and presentation attack detection. This approach has thee input components: two speaker embeddings (for enrollment and test each) and prediction of presentation attacks. Experiments are conducted using the ASVspoof 2017-v2 dataset, which includes official trials on the integration of speaker verification and presentation attack detection. The proposed back-end approach demonstrates a relative improvement of 21.77% in terms of the equal error rate for integrated trials compared to a conventional speaker verification system.

Keywords: speaker verification; presentation attack detection; deep neural networks

1. Introduction

Recent advances in deep neural networks (DNNs) have improved the performance of speaker verification (SV) systems, including short-duration and far-field scenarios [1–5]. However, SV systems are known to be vulnerable to various presentation attacks, such as replay attacks, voice conversion, and speech synthesis. These vulnerabilities have inspired research into presentation attack detection (PAD), which classifies given utterances as spoofed or not spoofed [6–8], where many DNN-based systems have achieved promising results [9–11].

Table 1 demonstrates the vulnerability of conventional SV systems when faced with presentation attacks. The performance is reported using the three types of equal error rates (EERs) described in Table 2 [12]. Table 2 shows the target and non-target trials for calculating the EER, which are represented by 1 and 0, respectively. Zero-effort (ZE)-EER describes the conventional SV performance without considering the presence of presentation attacks. PAD-EER denotes the EER for PAD which only considers whether an input is spoofed. Integrated speaker verification (ISV)-EER describes overall performance, considering both speaker identity and spoofing. We refer to "replay spoofing-aware SV" as an ISV task and report its performance using ISV-EER. Results show that the EER of SV degrades to 33.72% with replayed utterances; this fatal performance degradation supports the necessity of a

spoofing-aware ISV system. In this paper, PAD refers to replay attacks, because the ASVspoof2017 dataset only focuses on replay attack detection which is known to be the easiest yet effective attack. Three tasks are considered: SV, PAD, and ISV, and performance is evaluated using ZE-EER, PAD-EER, and ISV-EER.

Table 1. Difference in equal error rate (EER) according to the existence of replay non-target trials. Results demonstrate the vulnerability of speaker verification systems that are unaware of presentation attack detection (PAD).

	ZE-EER	PAD-EER	ISV-EER
SV baseline	9.58	33.72	19.98

Table 2. Three types of EERs reported in this paper: Enrollment utterance is always bona fide (i.e., genuine, not replayed). Target: enrollment and test utterances are uttered by an identical speaker and are bona fide; zero-effort (ZE) non-target: enrollment and test utterances are uttered by different speakers and are bona fide; Replay non-target: enrollment and test utterances are uttered by an identical speaker and test utterance is replay spoofed.

	Target	ZE Non-Target	Replay Non-Target
ZE-EER	1	0	
PAD-EER	1		0
ISV-EER	1	0	0

While a number of studies have worked to develop independent systems for SV and PAD, few have sought to integrate the SV and PAD systems [12–17]. More specifically, this handful of studies proposed approaches such as cascaded, parallel [12,13], and joint systems [14,16,17]. Most existing studies used common features to integrate the two tasks for system efficiency. Section 2 further takes up this existing body of work.

In this paper, we propose two spoofing-aware frameworks for the ISV task, illustrated in Figure 1. We use a light convolutional neural network (CNN) (LCNN) architecture [18] for both frameworks; this choice is based on its success in various PAD studies [11,19]. The first proposed framework expands existing work by proposing a monolithic end-to-end (E2E) architecture. More specifically, it conducts speaker identification (SID) and PAD to train a common feature using multi-task learning (MTL) [20]. Concurrently, it uses the embeddings to compose trials and conduct the ISV task. Using the sum of SID, PAD, and ISV losses, the entire DNN is jointly optimized. However, based on tendencies observed during internal experiments, we hypothesize that training a common feature for the ISV task may not be ideal because the properties required for each task differ: the PAD task representation uses device and channel information while SV needs to remove it (further discussed in Section 3).

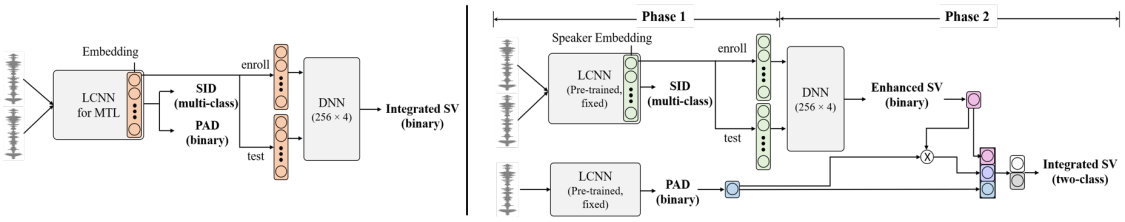


Figure 1. (**a**) An end-to-end architecture that trains embeddings (used for speaker identification (SID) and presentation attack detection (PAD)). LCNN and MTL refer to the light cnn and multi-task learning, concurrently; (**b**) a separate architecture that inputs speaker embeddings from SID and PAD results and outputs the Integrated result of speaker verification (SV) and PAD.

Based on our hypothesis, we propose a novel modular approach using a separate DNN. This approach inputs two speaker embeddings (for enrollment and test each) and a PAD prediction to make the ISV decision. It adopts a two-phase approach. In the first phase, the speaker identifier and PAD system are trained separately. In the second phase, speaker embeddings are extracted from a pretrained speaker identifier [21], and the embeddings and PAD prediction results are fed to a separate DNN module. Using this framework, we achieved a 21.77% relative improvement in terms of ISV-EER.(We use the trial in https://www.asvspoof.org/index2017.html for calculating ISV-EER.)

The contributions of this paper are as follows.

1. Propose a novel E2E framework that jointly optimizes SID, PAD, and the ISV task.
2. Experimentally validate the hypothesis that the discriminative information required for the SV and the PAD task may be distinct, requiring separate front-end modeling.
3. Propose a separate modular back-end DNN that takes speaker embeddings and PAD predictions as an input to make ISV decisions.

The remainder of the paper is organized as follows. Section 2 details related work on the integrated system of SV and PAD. Section 3 introduces the two proposed frameworks. Section 4 presents our experiments and results, and the paper is concluded in Section 5.

2. Related Work

In this section, we introduce the two studies most relevant to this study [12,16,17]. First, Todisco et al. [12] propose a separate modeling of two Gaussian back-end systems with a unified threshold for both SV and PAD tasks. Their study explores various acoustic features to find which ones best simultaneously suited both tasks. As organizers of the ASVspoof challenges, official trials for the ISV task are released in this study. For our purposes, it is important to highlight that these trials include both ZE and replayed non-target, which we use throughout this paper. However, Todisco et al. [12] reported the average of two EERs—ZE-EER and PAD-EER—because they separately modeled two Gaussian mixture models for each task.

Li et al. [16,17] extended Todisco et al.'s work [12] by proposing an integrated ISV system; this study is the first that reports an ISV-EER. More specifically, they propose a three-phase training framework for extracting an embedding for the ISV task, followed by a probabilistic linear discriminant analysis (PLDA) back-end. In the first phase, a MTL [20] framework is employed to train a common embedding for both SV and PAD tasks. In the second and third phases, the embedding is adapted to fit the ISV task. However, because the DNN is adapted in the third phase to fit the enrollment speakers, it has limitations for real-world scenarios. In addition, because the performance is reported does not exploit organizer's official trials, it is difficult to compare the performance with the literature.

In this paper, we first propose an E2E framework, illustrated in Figure 1a, that extends the work of Li et al. [16,17] in two aspects: First, we adopt a single phase training approach by using three loss functions for SID, PAD, and ISV. Second, our framework directly outputs a spoofing-aware score without using a separate back-end system.

3. Integrated Speaker Verification

In this section, we describe the two proposed frameworks for conducting speaker verification that are aware of presentation attacks, as shown in Figure 1.

3.1. End-to-End Monolithic Approach

We first propose an E2E monolithic approach. This architecture simultaneously trains all components, including SID, PAD, and ISV, using a common feature, as illustrated in Figure 1a. The loss function for training the proposed E2E architecture comprises three components: a categorical cross-entropy (CCE) loss for SID, a binary cross-entropy (BCE) loss for PAD, and a two-class BCE loss for ISV. When a mini-batch is input for training, the proposed system first conducts SID and PAD with

an MTL framework. Then, it composes a number of trials. A trial consists of two embeddings: one for enroll and the other for test. The ISV prediction is made by feed-forwarding the two embeddings through a few fully-connected layers. The entire DNN is jointly optimized using the sum of three loss functions. The objective function *Loss* is defined as follows,

$$Loss = Loss_{SID} + Loss_{PAD} + Loss_{ISV} \quad (1)$$

where $Loss_{SID}$ refers to the CCE loss for SID, $Loss_{PAD}$ is the BCE loss for PAD, and $Loss_{ISV}$ denotes the CCE loss for ISV.

However, we find consistent tendencies that make it difficult to extract a common representation, i.e., feature, for performing both SV and PAD tasks through experiments. Therefore, we hypothesize that, although SV and PAD tasks are closely related in the scenario, the discriminative information required for each task collides. Speaker embeddings for the SV task requires robustness to device and channel difference; meanwhile, representation for the PAD task uses such information [22]. Additionally, both bona fide and replayed utterances include the same speaker information, making it a less discriminative factor for the PAD task; meanwhile, it is key information for the SV task. The study of Sahidullah et al. [13] supports our hypothesis, which states that the SV and PAD tasks should exist independently. To validate our hypothesis, we conduct experiments using separately trained SV and PAD systems and MTL-based systems. We further detail these elements in Section 4.3.

3.2. Back-End Modular Approach

We also propose a novel modular approach using a separate DNN that takes speaker embeddings and PAD predictions as input to make an ISV decision. Figure 1b illustrates our second proposed system. Based on the hypothesis addressed in the previous subsection, we design an integrated system using a two-phase approach. In the first phase, we separately train an SID system to extract speaker embeddings from the last hidden layer and a PAD system to extract a spoofing prediction. Then, we train the ISV system by using two speaker embeddings (one for enroll and the other for test) extracted from the SID system as a pair and a PAD label as an input. This system has an output layer with two nodes: the first node indicates "acceptance", and the second node indicates "rejection" for both ZE and replay trials.

In Figure 1b, the part trained in phase 2 is the proposed back-end ISV system. It takes two speaker embeddings and multiplication of the two embeddings as input and a module of four fully-connected layers outputs a scalar that indicates whether they were uttered by the same speaker. The fully-connected layers comprise 256 nodes each and an output layer comprises one node with a sigmoid function.

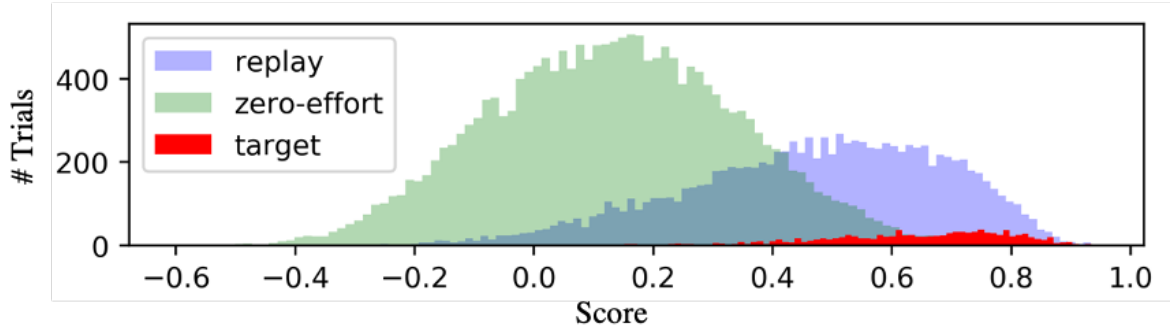
Next, the SV and PAD prediction results and their multiplication are fed to a fully-connected layer to make the final decision. In an ideal scenario, the multiplication of the SV result and PAD prediction would indicate 1 when both SV and PAD are positive, and 0 otherwise; we assume this multiplication would additionally inform the final decision. The objective function $Loss_{int}$ for the back-end modular approach comprises loss for the SV task and the loss for the final decision, defined as

$$Loss_{int} = \alpha \cdot Loss_{SV} + Loss_{ISV} \quad (2)$$

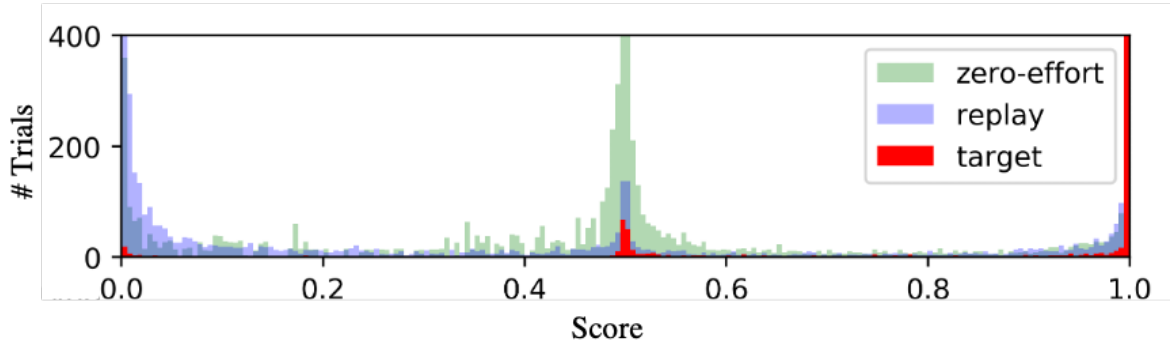
where $Loss_{SV}$ and $Loss_{ISV}$ refer to the BCE loss of the SV task and the CCE loss of the ISV task, respectively, and α signifies the weight for the SV loss. We note that training the proposed back-end DNN with only $Loss_{ISV}$ results in overfitting.

Based on a number of experiments that we omit here for the sake of brevity, we find two key components that make our proposed back-end DNN framework successful: First, we aim to model ZE and replayed trials into separate score distributions. Figure 2a,b, respectively, illustrates the score distributions of the evaluation trials of the SV baseline and the proposed modular back-end DNN. In Figure 2a, the score refers to the cosine similarity of the two embeddings. Here, the score

distribution of replay non-target trials severely overlaps with that of target trials. As the existing speaker verification system does not consider the presentation attack detection, replay non-target and target trials can only be determined as targets because the replayed utterances and the bona fide utterances share the same speaker information. In our analysis, this resulted from embeddings that only considered speaker information in which replayed and bona fide utterances coincided. In various experiments, it is impossible to model both replay and ZE non-target trials into the same score distribution. When one kind of non-target trial was successfully modeled, the other resulted in a distribution similar to uniform. Therefore, we aim to separate two non-target score distributions, specifically by modeling the score distribution of ZE non-target to have a mean of 0.5 and replay the non-target to have a zero mean. To do so, we sequentially apply rectified linear unit (ReLU) and sigmoid activation functions to the output of SV before the last hidden layer for ISV. Figure 2b demonstrates the score distribution of the proposed method. The results demonstrate that three types of evaluation trials were modeled as intended (i.e., well generalized) in the case of evaluation trials, though these trials comprised unknown speakers and replay conditions.



(**a**)



(**b**)

Figure 2. Histograms of score distribution on the evaluation trials. (**a**) Speaker verification (SV) baseline, where score is calculated using cosine similarity of two speaker embeddings. (**b**) The proposed modular system, where three types of trials have three different distributions.

Second, we use actual PAD labels instead of PAD predictions of the spoofing DNN in the training phase. This is based on empirical comparisons in which the use of PAD predictions in the training phase worsened the performance. In our analysis, using PAD labels in the training phase was more helpful because even a small number of misclassified utterances among PAD predictions can interrupt the training of the proposed DNN. Notably, we empirically observed model collapse when training the proposed modular DNN using PAD predictions.

4. Experiments and Results

4.1. Dataset

All experiments in this study were conducted using the ASVspoof2017-v2 dataset [23], because the official trials do not exist for integrated systems for the ASVspoof2019 dataset. Therefore, we evaluated the proposed integrated system by utilizing the official trials reported in [12] on the ASVspoof2017-v2 dataset. We used training and development sets to train all systems comprising 2267 bona fide and 2457 replay spoofed utterances from 18 speakers. To evaluate speaker verification and presentation attack detection performances, we measured the ZE-EER and the PAD-EER using the ASVspoof2017 joint PAD+SV evaluation trial. This trial comprised 1106 target, 18,624 ZE, and 10,878 replayed trials. We used target and ZE for ZE-EER and target and replayed for PAD-EER evaluations.

4.2. Experimental Configurations

We used PyTorch, a Python deep learning library, for all experiments. For all DNNs, we input 64-dimensional Mel-filterbank features, with utterance-level mean normalization following [22]. We applied weight decay with $\lambda = 1e^{-4}$, and optimized with an AMSGrad optimizer [24].

Regarding our use of ASVspoof2017-v2, we found that relatively thin LCNN structures were helpful for performance improvement; this may have been a result of the small size of the dataset. In addition, we also found that minute changes to the DNN greatly influence the performance because of the small data scale; therefore, a relatively thin structure remained particularly helpful for performance improvement. To derive a value between 0 and 1 for the PAD task, we used a network architecture identical to that of [11] but replaced the angular margin softmax activation [25] with a sigmoid function. We also modified the architecture for the SV task based on [11]. Speaker embeddings had a dimensionality of 1024.

4.3. Results Analysis

Table 3 describes the results of the proposed E2E framework with a monolithic approach. System #1 refers to the proposed architecture that jointly optimizes SID, PAD, and ISV loss (see Figure 1a). System #2-SE is the result of applying squeeze-excitation (SE) [26] based on its recent application to PAD [9]. System #3 describes the result of assigning three max feature map (MFM) blocks [18] for SID as well as for PAD after the first three MFM blocks. Because most of the system's performance measures deteriorated compared to the SV baseline, we concluded that the monolithic E2E approach was not ideal for the ISV task. While the results of the experiments were different from what we expected, they nevertheless served as a springboard for establishing a new hypothesis.

Table 3. Results of various architectures using the proposed monolithic E2E framework for the ISV task. Numbers in bold represent the best results. Numbers in bold represent the best results.

System	ZE-EER (SV)	PAD-EER	ISV-EER
#1	**18.52**	**15.73**	18.44
#2-SE	18.99	15.90	**17.90**
#3-split	19.43	37.31	26.40

Table 4 addresses the validation of our hypothesis in Section 3 that the discriminative information for the SV and the PAD task are distinct based on the results of Table 3. To validate our hypothesis, we trained our SV and PAD baselines with and without additional loss for extracting common embeddings. Here, the first and third rows refer to the SV and PAD baselines and the second and fourth rows refer to the usage of the MTL framework. The results demonstrated that, in both baselines, additionally adopting another loss function degraded performance.

Table 4. Experimental results showing that the required discriminative information differs for SV and PAD (Sid: speaker identification; PAD: presentation attack detection; Int: integrated speaker verification). Numbers in bold represent the best results.

Train Loss	DNN Arch	ZE-EER (SV)	PAD-EER
Sid	SV	**9.58**	-
Sid+PAD	SV	17.53	13.69
PAD	PAD	-	**10.60**
PAD+Sid	PAD	19.16	12.17

Table 5 summarizes the results of performance improvement across various attempts to improve the performance of the proposed method in the back-end modular approach. The comparison of Systems #4 and #5 shows the effectiveness of using multiplication of the SV result and PAD prediction for the ISV task. System #6 refers to the result of setting weights to the SV task in the training phase where we set the α to 20. System #7 shows the result of reducing the number of nodes per hidden layer.

Table 5. Results of the proposed modular approach for the ISV task. Numbers in bold represent the best results.

System	ZE-EER (SV)	PAD-EER	ISV-EER
#4-w/o mul	20.52	19.77	20.48
#5-w mul	15.59	18.06	16.66
#6-loss weight	15.22	**14.55**	15.91
#7-DNN arch	**14.32**	15.46	**15.63**

Finally, Table 6 compares our proposed modular approach with the SV baseline and existing work [12] using official trials. The results demonstrated that the proposed approach stabilizes unbalanced performance between ZE-EER and PAD-EER. Compared with the SV baseline, which does not consider PAD attacks, we achieved a relative improvement of 21.77%. It is important to note here that we were unable to compare the ISV-EER with that of Todisco et al. [12], although it is the only study that reported performance using official trials. Because it proposed a unified threshold for conducting SV and PAD tasks, ISV-EER results using the full trial do not exist.

Table 6. Comparison of the SV baseline, our proposed modular deep neural network (DNN), and other work using the official trials for the ISV task. Numbers in bold represent the best results.

	ZE-EER	PAD-EER	ISV-EER
SV Baseline	9.58	33.72	19.98
#7-Ours	14.32	**15.46**	**15.63**

5. Conclusions

In this paper, we investigated the integration of speaker verification and presentation attack detection. We proposed two methods for their integration: an E2E monolithic approach and a back-end modular approach. The proposed E2E approach simultaneously trains SID, PAD, and ISV using a common feature. The experimental results of the E2E approach led us to hypothesize that the discriminative information for SID and PAD differs. Based on our hypothesis, we proposed a framework using a separate back-end DNN that takes speaker embedding and a PAD prediction extracted from pretrained SV and PAD systems as input. The effectiveness of our proposed systems was verified using official trials for the ISV task, where we achieved an EER of 15.63%. It is expected that the proposed method will continue to enhance performance when improved speaker embeddings and PAD prediction are input.

Author Contributions: Conceptualization, investigation, and writing—original draft preparation and editing, H.-j.S. and J.-w.J.; writing—review and editing, J.-h.K.; supervision, writing—review and editing, H.-j.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science, ICT and Future Planning, Grant number PA-J000001-2017-101.

Acknowledgments: This research was supported by Projects for Research and Development of Police Science and Technology under the Center for Research and Development of Police Science and Technology and the Korean National Police Agency funded by the Ministry of Science, ICT and Future Planning (Grant No. PA-J000001-2017-101).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bhattacharya, G.; Alam, J.; Kenny, P. Deep speaker recognition: Modular or monolithic? In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1143–1147.
2. Jung, J.W.; Heo, H.S.; Kim, J.H.; Shim, H.J.; Yu, H.J. RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1268–1272.
3. Tawara, N.; Ogawa, A.; Iwata, T.; Delcroix, M.; Ogawa, T. Frame-Level Phoneme-Invariant Speaker Embedding for Text-Independent Speaker Recognition on Extremely Short Utterances. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6799–6803.
4. Jin, Q.; Schultz, T.; Waibel, A. Far-field speaker recognition. *IEEE Trans. Audio Speech Lang Process* **2007**, *15*, 2023–2032. [CrossRef]
5. Jung, J.; Heo, H.; Shim, H.; Yu, H. Short Utterance Compensation in Speaker Verification via Cosine-Based Teacher-Student Learning of Speaker Embeddings. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 335–341.
6. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
7. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2–6.
8. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. *arXiv* **2019**, arXiv:1904.05441
9. Lai, C.I.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-Spoofing with squeeze-excitation and residual networks. *arXiv* **2019**, arXiv:1904.01120.
10. Jung, J.W.; Shim, H.J.; Heo, H.S.; Yu, H.J. Replay Attack Detection with Complementary High-Resolution Information Using End-to-End DNN for the ASVspoof 2019 Challenge. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1083–1087.
11. Lavrentyeva, G.; Novoselov, S.; Tseren, A.; Volkova, M.; Gorlanov, A.; Kozlov, A. STC antispoofing systems for the ASVSpoof2019 challenge. *arXiv* **2019**, arXiv:1904.05576.
12. Todisco, M.; Delgado, H.; Lee, K.A.; Sahidullah, M.; Evans, N.; Kinnunen, T.; Yamagishi, J. Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 77–81.
13. Sahidullah, M.; Delgado, H.; Todisco, M.; Yu, H.; Kinnunen, T.; Evans, N.; Tan, Z.H. Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 1700–1704.
14. Sizov, A.; Khoury, E.; Kinnunen, T.; Wu, Z.; Marcel, S. Joint Speaker Verification and Antispoofing in the *i*-Vector Space. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 821–832. [CrossRef]

15. Dhanush, B.; Suparna, S.; Aarthy, R.; Likhita, C.; Shashank, D.; Harish, H.; Ganapathy, S. Factor analysis methods for joint speaker verification and spoof detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5385–5389.
16. Li, J.; Sun, M.; Zhang, X. Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1517–1522.
17. Li, J.; Sun, M.; Zhang, X.; Wang, Y. Joint Decision of Anti-Spoofing and Automatic Speaker Verification by Multi-Task Learning With Contrastive Loss. *IEEE Access* **2020**, *8*, 7907–7915. [CrossRef]
18. Wu, X.; He, R.; Sun, Z.; Tan, T. A light CNN for deep face representation with noisy labels. *arXiv* **2015**, arXiv:1511.02683.
19. Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio Replay Attack Detection with Deep Learning Frameworks. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 82–86.
20. Caruana, R.A. *Multitask Learning: A Knowledge-Based Source of Inductive Bias*; Learning to Learn; Springer: Berlin/Heidelberg, Germany, 1998.
21. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056.
22. Shim, H.J.; Jung, J.W.; Heo, H.S.; Yoon, S.H.; Yu, H.J. Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In Proceedings of the Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taichung, Taiwan, 30 November–2 December 2018; pp. 172–176.
23. Delgado, H.; Todisco, M.; Sahidullah, M.; Evans, N.; Kinnunen, T.; Lee, K.A.; Yamagishi, J. ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements. In Proceedings of the Odyssey 2018 The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018; pp. 296–303.
24. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237.
25. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE cOnference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.