Beyond Silence:
Bias Analysis through Loss and Asymmetric Approach in Audio Anti-Spoofing

Hye-jin Shim[1], *Md Sahidullah*[2], *Jee-weon Jung*[1], *Shinji Watanabe*[1], *Tomi Kinnunen*[3]

[1]Carnegie Mellon University, USA
[2]TCG CREST, India
[3]University of Eastern Finland, Finland

shimhz6.6@gmail.com

Abstract

Current trends in audio anti-spoofing detection research strive to improve models' ability to generalize across unseen attacks by learning to identify a variety of spoofing artifacts. This emphasis has primarily focused on the spoof class. Recently, several studies have noted that the distribution of silence differs between the two classes, which can serve as a shortcut. In this paper, we extend class-wise interpretations beyond silence. We employ loss analysis and asymmetric methodologies to move away from traditional attack-focused and result-oriented evaluations towards a deeper examination of model behaviors. Our investigations highlight the significant differences in training dynamics between the two classes, emphasizing the need for future research to focus on robust modeling of the bonafide class.

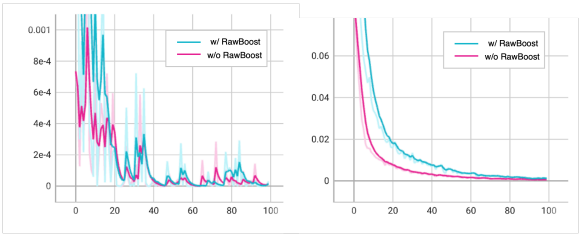
Index Terms: anti-spoofing, deepfake detection, spoofing detection, shortcut learning, ASVspoof

1. Introduction

Recent progress in voice conversion (VC) and text-to-speech (TTS) technologies have intensified concerns regarding their potential for malicious use, emphasizing the critical role of audio anti-spoofing systems. Audio spoofing detection systems, as binary classifiers, distinguish genuine human speech (*bonafide*) from artificially generated (*spoofed*) speech. The main direction in this field is toward advancing the systems' ability to generalize across unseen spoofing attacks by focusing on learning diverse spoofing artifacts. To this end, a pivotal shift from traditional hand-crafted features [1, 2] to data-driven approaches [3–5] and data augmentation [6, 7] have been pursued. Research on diversifying training data [8–10], incorporating domain adaptation strategies [11, 12], and leveraging large-scale pre-trained models [13–15] also follow this trend.

Despite these efforts, the evolution of TTS and VC systems, especially those utilizing state-of-the-art methods like diffusion, are likely to involve fewer spoofing artifacts and create subtle variations in them without hard effort. This suggests that learning diverse spoofing artifacts does not guarantee the detection of emerging unknown attacks. To tackle this, several studies have explored methods for modeling robust bonafide features and distinguishing them from spoof features within the latent space [16–19]. Class-wise analyses have sought to determine the differences between bonafide and spoof classes as detailed in [18, 20, 21] and their findings primarily converge on the impact of 'silence', which has a spurious correlation with spoofing detection.

These kinds of external factors that could unintentionally lead the biased model predictions are known as shortcuts [22], recently getting a lot of attention in the broader deep learning literature. These shortcuts challenge the ability to determine if a model genuinely distinguishes between classes or relies on irrelevant cues associated with class labels. In the context of audio anti-spoofing, 'silence' is a well-known *data bias*; for instance, the model trained with



(a) Bonafide class loss (b) Spoof class loss

Figure 1: *Comparison of training loss. The left and right figures illustrate bonafide and spoof classes. x-axis and y-axis indicate training epochs and loss magnitude. Regardless of the implementation of data augmentation, the two class losses differ on a large scale.*

ASVspoof2017 dataset [23] is vulnerable to silence [24] and the ASVspoof2019 dataset [25] exhibits an unequal distribution of silence between spoofed and bonafide samples [26,27]. The impact of silence varies not only with the data source but also with its distinct types/patterns introduced during data collection. Efforts to understand and mitigate the influence of silence, including the analysis of silence trimming and its effects, have been undertaken [18,26]. Other than silence, unveiled factors also have been studied in [28], recently.

While in-depth analyses in audio anti-spoofing have significantly advanced the field and provided invaluable insights, several areas remain for further explorations: (i) the majority of studies have focused on spoof class, specifically per-attack interpretations; (ii) analyses on each class exist, yet the emphasis has primarily been on silence, except for [28] exploring other data bias factors; (iii) existing analyses predominantly rely on evaluating outcomes, such as performance metrics and score distributions, without a thorough examination of the internal workings of models during training. Our research diverges from existing studies by delving into the training process through loss analysis and adopting a novel asymmetric intervention on each class and phase(train and test) to understand the respective effects.

Our analysis starts with examining the loss curves for bonafide and spoof classes separately, both with and without the application of RawBoost [7] data augmentation, as illustrated in Figure 1. Note that illustrated loss values are raw values. For the model update, a loss weight of 0.9 (bonafide) and 0.1 (spoof) was employed, considering the imbalanced number of samples in each class. Interestingly, the results indicate that the loss associated with the bonafide class is significantly lower than that for the spoof class. Even if one considers the loss weights in the training phase, a substantial gap still remains between the two classes. These results could signify that the bonafide class is inherently easier to train than the spoof class. Alternatively, this might suggest that modeling the bona fide class is not trivial; however, there exists a shortcut that significantly reduces the training loss. We deploy various loss functions to reveal the meaning behind the low magnitude of the bonafide loss, between the two scenarios mentioned. In particular, the objective functions

that can consider the difficulty of samples enable us to understand how the model deals with each class based on our findings.

Furthermore, our approach involves training and testing the model with an asymmetric intervention to assess the bonafide and spoof classes separately. Unlike the methodology in [28], which applies interventions across both classes and phases for a comprehensive model-level interpretation, our strategy focuses interventions on one side only. This allows a more precise understanding of how different phases and class-specific traits affect model performance. Additionally, to mitigate potential concerns regarding the influence of silence, we also conduct our analysis combined with silence trimming as well, demonstrating that our findings are not biased by the presence of silence. Our findings pave the way for new directions of future research, shifting focus from the currently predominant studies centered around the spoof class.

2. Method

In this section, we introduce two primary methodologies that facilitate our in-depth investigations of anti-spoofing systems: (i) a loss-based analysis that contrasts objective functions, distinguishing between those that prioritize either easier or more challenging samples, and (ii) an asymmetric intervention analysis that examines the impact of intervention on phases (train or test) and classes (bonafide or spoof).

2.1. Loss based analysis

Previous analyses of silence within audio anti-spoofing contexts have paved the way for separate examinations of bonafide and spoof classes. This study extends these insights by exploring the various objective functions that can help us understand the model's behavior in each class; loss analysis is widely adopted in machine learning for understanding the model behavior during the training [29]. Furthermore, the loss has been directly employed in the image domain to infer data bias in [30–32] by analyzing its correlation with sample difficulty and data bias.

In this work, we conduct two types of analysis using loss functions. Initially, we calculate the loss for bonafide and spoof classes separately during training to observe their convergence patterns and magnitude differences. Significant differences in these areas may indicate model bias, as shown in Figure 1. Subsequently, we differentiate between two categories of loss functions for assessing model behavior: prioritizing hard *or* easy samples. Among diverse loss functions utilized in hard negative mining, we deliberately choose a few that dynamically modulate each sample's impact on the overall loss, rather than explicitly selecting samples. This strategy enables an in-depth exploration of the model's inherent reactions to varying sample types.

FocalLoss [33] prioritizes hard samples by modifying the categorical cross-entropy loss. It amplifies the loss for samples with inaccurate model predictions or where the model exhibits uncertainty (i.e., low predicted probability for the correct class), thus prioritizing hard or incorrectly classified samples without resorting to specific sample selection strategies.

SuperLoss [34] assigns a weight to each sample based on a moving average of its past losses, targeting potentially noisy or outlier samples. This method progressively focuses on challenging or ambiguous samples.

CurricularFace [35] emphasizes more challenging samples by increasing their relative loss compared to easier ones using class-specific margins. It adjusts the target margins for classes, making it easier or harder for the model to classify them correctly as training progresses.

Generalized cross entropy (GCE) [36], in contrast, focuses on *easier* samples. It enhances the penalty for misclassifying classes with lower probabilities, thereby prioritizing minority classes. This method can be particularly useful in audio anti-spoofing to direct the model's attention towards bonafide samples. In addition, GCE is employed in model debiasing research to concentrate on biased samples, aligning with empirical observations that such samples exhibit lower loss values during training.

Table 1: *Five experimental configurations using eight subsets of a dataset.* ***O*** *is the original (unintervened) configuration. All other four sets have one particular subset intervened.*

Intervened phase	Configuration	Train set	Test set
-	**O**	$\mathcal{D}_{\text{trn, bona}} \cup \mathcal{D}_{\text{trn,spf}}$	$\mathcal{D}_{\text{test, bona}} \cup \mathcal{D}_{\text{test, spf}}$
Train	**Tr_B**	$\mathcal{D}^{\textbf{intervened}}_{\text{trn, bona}} \cup \mathcal{D}_{\text{trn,spf}}$	$\mathcal{D}_{\text{test, bona}} \cup \mathcal{D}_{\text{test, spf}}$
	Tr_S	$\mathcal{D}_{\text{trn, bona}} \cup \mathcal{D}^{\textbf{intervened}}_{\text{trn,spf}}$	$\mathcal{D}_{\text{test, bona}} \cup \mathcal{D}_{\text{test, spf}}$
Test	**Te_B**	$\mathcal{D}_{\text{trn, bona}} \cup \mathcal{D}_{\text{trn,spf}}$	$\mathcal{D}^{\textbf{intervened}}_{\text{test, bona}} \cup \mathcal{D}_{\text{test, spf}}$
	Te_S	$\mathcal{D}_{\text{trn, bona}} \cup \mathcal{D}_{\text{trn,spf}}$	$\mathcal{D}_{\text{test, bona}} \cup \mathcal{D}^{\textbf{intervened}}_{\text{test, spf}}$

2.2. Asymmetric intervention analysis

Our analysis follows an interventional approach proposed recently in [28] to reveal shortcuts in speech anti-spoofing beyond silence. The essence of this approach is to intentionally modify (intervene) an existing dataset to provoke the classifier to rely on shortcuts. We start by reviewing the approach and then explain how we modify it in this study.

To explore potential shortcuts in audio anti-spoofing, [28] employed standard audio manipulations (e.g., MP3 compression or additive noise) with randomized parameters applied to datasets to create artificial statistical associations between the audio and the class label. For example, applying MP3 compression to only bonafide data in both training and test data while leaving spoof data unaltered resulted in equal error rates (EERs) of 0%. Conversely, applying MP3 compression exclusively to spoofed test data while keeping all other conditions the same led to an opposite outcome (EER $> 99\%$), illustrating a complete label flip. Such extreme interventions reveal the vulnerability of the spoofing detection system against potential bias factors unrelated to spoofing artifacts residing within the data.

The previous approach concurrently applied interventions during both training and test phases to understand the overall model's behaviors under different interventions, producing the boundary cases of 'near-perfect' outcomes and 'worse than coin flip' (label flip) results. In contrast, our investigation takes a class-wise interpretive approach based on the understanding that each class responds differently. Consequently, we strategically apply interventions at only one phase or one class at a time, leveraging two binary dimensions: phase (training or testing) and class (bonafide or spoof). Our approach enables (i) to reveal the interventions associated with each class separately, (ii) to evaluate the robustness of the model's representation for each class separately, and (iii) to compare the robustness of class modeling by observing the outcomes of intervention.

In particular, our methodology of employing class- and phase-wise intervention in either class in either phase results in four distinct intervention configurations. Formally, let $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathbb{X}, y_i \in \mathbb{Y}, \text{for } i = 1,2,...,N$, where $\mathcal{D}$ represents the dataset, (x_i, y_i) denotes an individual sample, and N is the number of samples. Typically, the corpus is divided into two subsets: the training ($\mathcal{D}_{\text{trn}}$) and the test ($\mathcal{D}_{\text{test}}$) sets. We introduce another categorization based on class, dividing the dataset into four subsets: $\mathcal{D}_{\text{trn, bona}}$, $\mathcal{D}_{\text{trn, spf}}$, $\mathcal{D}_{\text{test, bona}}$, $\mathcal{D}_{\text{test, spf}}$. Furthermore, we generate four additional subsets with applied interventions, where the labels for each sample remain: $\mathcal{D}^{\textbf{intervened}}_{\text{trn, bona}}$, $\mathcal{D}^{\textbf{intervened}}_{\text{trn, spf}}$, $\mathcal{D}^{\textbf{intervened}}_{\text{test, bona}}$, $\mathcal{D}^{\textbf{intervened}}_{\text{test, spf}}$. Note that the intervention does not affect the ground-truth label $\mathbb{Y}$. Composing the eight total subsets, we present five experimental configurations in Table 1: **O**, **Tr_B**, **Tr_S**, **Te_B**, and **Te_S**. Therefore, the comparison of **B** and **S**, modification on the class

Table 2: *Performance comparison on different loss functions. Apart from original categorical cross entropy (CCE), FocalLoss, SuperLoss, and CurricularFace are designed to concentrate on hard samples based on empirical loss while generalized cross entropy (GCE) focuses on easier samples. All evaluations are conducted on the ASVspoof2019 LA dataset.*

Loss function	EER (%)
Original (CCE)	1.39
FocalLoss [33]	1.67
SuperLoss [34]	1.45
CurricularFace [35]	1.53
GCE [36]	**1.35**

Table 3: *Performance comparison in EER when the model is **tested** on asymmetric ways. All models are trained on the original training dataset but evaluated in different configurations. Here, the ratio indicates a relative change of **Te_B** compared to that of **Te_S**.*

System	**O**	**Te_S**	**Te_B**	Ratio
MP3				
AASIST	0.83	0.77	9.07	137.33
AASIST-L	1.39	1.21	8.78	41.06
AASIST-L w/ RawBoost	1.59	1.62	3.96	79.00
AASIST-L w/ MDL	0.99	1.29	3.63	8.80
Noise				
AASIST	0.83	0.39	31.08	68.75
AASIST-L	1.39	0.65	35.50	46.09
AASIST-L w/ RawBoost	1.59	1.06	13.27	22.04
AASIST-L w/ MDL	0.99	0.69	8.24	24.17
Loudness				
AASIST	0.83	1.18	5.33	12.86
AASIST-L	1.39	2.23	6.73	6.36
AASIST-L w/ RawBoost	1.59	1.52	2.38	11.29
AASIST-L w/ MDL	0.99	1.44	4.09	6.89

side, enables us to analyze the class-wise effect of interventions. Comparison of **Tr** and **Te**, on the other hand, helps to understand the class-wise difference and how the model operates differently depending on whether the model learns such intervened condition.

3. Experimental setup

3.1. Datasets

ASVspoof2019 [25] is a dataset for logical access (LA) scenario of audio anti-spoofing and it includes 6 and 13 types of spoofing attacks in train/dev and test. The number of utterances for train/dev and test is 50,224 and 71,237, respectively.

ASVspoof2021 [37] includes the latest spoofing attacks, which consists of different test sets, each for LA and deepfake (DF) scenarios. The LA and DF subsets include diverse synthetic techniques and audio compressions, respectively.

3.2. Implementation details

Model architectures used in this paper are AASIST, AASIST-L [38] models with data augmentation and multi-dataset co-training. Those two models are state-of-the-art models that directly operate on raw waveform and they only differ in the number of parameters.

Intervention types are selected among five different interventions in [28]. We employ three interventions: *MP3 compression*, *additive white noise*, and *loudness normalization*. Those interventions are considered since MP3 compression and white noise are discovered as the most influential ones, while loudness normalization is the least effective one.

Data augmentation is implemented to check the difference when we employ the model considered more robust. We utilize RawBoost [7] which includes three different augmentation techniques: linear and non-linear convolutive noise, multi-band filters, and Hammerstein systems [39]. We deploy three of them simultaneously as it showed the best result in [7].

Multi-dataset trained model with sharpness optimization [10] is utilized to investigate the robust model similar to data augmentation. To further the enhance generalization capability, the model is trained using multiple datasets at the same time and optimized by sharpness-aware optimization [40]. We select the model trained by both ASVspoof2015 and ASVspoof2019 LA with adaptive sharpness-aware minimization (ASAM) [41] as it showed the best performance in ASVspoof2019 LA evaluation by 0.99% of EER.

Silence trimming is implemented to mitigate the influence of silence that might distort the imbalance results. Our silence trimming algorithm works as follows. First, we detect the speech frames using a simple energy-based algorithm as described in [42]. We have used a frame size of 25 ms with an 8ms shift. Then we remove the silences where the silence length is more than 50 ms.

4. Results and Analysis

4.1. Comparison of class-wise loss and different objective functions

As previously introduced in Section 1 with Figure 1, the observed loss curve reveals an unexpected pattern, especially considering the predominance of spoof data in the training set; in general, more frequently represented spoof classes are expected to have lower losses, consistent with the general bias of neural networks to focus on more frequently represented classes. The results indicate that effective training of the bonafide class is hindered not only by its fewer amount of samples in the training dataset but also by its comparatively lower loss. Consequently, the neural network would prioritize minimizing the spoof class's loss in which the loss scale is much higher, leading to an inherent bias, which is unintended. This research marks the first to identify this particular bias within the context of audio spoofing detection, shedding light on the challenges posed by class imbalance and its impact on model training dynamics.

Building upon these insights, we further examine the potential class bias in model training towards the spoof class using diverse loss functions outlined in Section 2.1. Our comparison, shown in Table 2, involves four different loss functions – three (FocalLoss [33], SuperLoss [34], and CurricularFace [35]) designed for challenging (spoof) samples and one (GCE [36] for easier (bonafide) samples. The results reveal a decline in performance with the three loss functions aimed at the spoof class, while the GCE, which prefers the bonafide class, slightly improves model performance. Despite adjustments using tunable parameters in FocalLoss, SuperLoss, and CurricularFace, all outcomes showed no significant enhancement[1]. Results highlight the training bias towards the spoof class. This pattern of bias and its impact on performance encourages us to call for a significant shift in focus for future audio anti-spoofing efforts, emphasizing **the importance of robust modeling the bonafide class** over solely concentrating on detecting spoofing artifacts. In addition, the field of anomaly sound detection [43] further demonstrates the advantages of focusing on bonafide modeling. State-of-the-art systems prioritize modeling the normal sound class and effectively identify deviations significantly distant from these norms as anomalies.

[1] Presented performances are the best results among our experiments.

Table 4: *Performance comparison in EER when the model is **tested** on asymmetric ways with silence trimming. Silence trimming is applied in both the training and the test phases; we additionally report the results when the silence is only removed from either phase.*

System	Silence trimming	**O**	MP3			Noise			Loudness		
			Te_S	**Te_B**	Ratio	**Te_S**	**Te_B**	Ratio	**Te_S**	**Te_B**	Ratio
AASIST-L	-	1.39	1.21	8.78	41.06	0.65	35.50	46.09	2.23	6.73	6.36
AASIST-L w/ RawBoost	-	1.59	1.62	3.96	79.00	1.06	13.27	22.04	1.52	2.38	11.29
AASIST-L	train	35.05	28.36	39.32	0.64	47.09	28.14	0.57	27.44	47.28	1.61
	test	25.14	20.07	37.7	2.48	8.27	59.04	2.01	13.92	40.83	1.40
	train&test	18.65	17.52	30.65	10.62	17.24	34.49	11.23	14.79	26.99	2.16
AASIST-L w/ RawBoost	train	45.52	40.65	48.47	0.61	66.93	31.14	0.47	36.10	55.35	1.04
	test	27.79	29.5	30	1.29	30.74	26.85	0.32	21.82	37.05	1.55
	train&test	19.73	17.56	31.69	5.51	21.07	32.35	9.42	12.75	29.73	1.43

Table 5: *Performance comparison in EER when the model **trained** on asymmetric ways. All evaluations are conducted using original test set without interventions.*

System	Intervention type	Config.	2019 LA	2021 LA	2021 DF
Original	-	-	1.39	12.18	21.8
RawBoost	-	-	1.59	6.42	17.48
RawBoost	MP3	**Tr_B**	16.21	15.45	30.86
		Tr_S	3.68	17.81	23.1
	Noise	**Tr_B**	71.02	62.45	62.66
		Tr_S	15.48	24.12	30.06
	Loudness	**Tr_B**	13.05	10.66	18.7
		Tr_S	1.82	9.49	17.97

4.2. Asymmetric test results

Table 3 presents the results of applying asymmetric interventions to a specific class during the test phase. These results are analyzed from two angles: (i) evaluating the intervention's impact by comparing the original results (**O**) with those after interventions on the spoof (**Te_S**) or bonafide (**Te_B**) classes (columns 2 and 3), and (ii) examining how interventions differently affect the bonafide class compared to the spoof class, as indicated in column 4. From the first angle, enhancements in performance post-intervention hint at possible model biases due to shortcuts. Conversely, declines indicate the intervention's irrelevance to the target class or might reflect domain differences or unknown factors. The ratio in column 4, calculated as the relative performance impact $(|\mathbf{O}-\mathbf{Te\_B}|/\mathbf{O})/(|\mathbf{O}-\mathbf{Te\_S}|/\mathbf{O})$, offers deeper insight into which class the intervention impacts more. Here, ratio >1 refers to a greater influence on the bonafide class, while ratio <1 implies spoof class is more affected.

Our findings highlight two main observations that align with our analysis of loss functions. First, interventions are prone to improve **Te_S** (e.g., AASIST with MP3 or AASIST-L with Noise on **Te_S**), while consistently lowering **Te_B** performance. This implies that factors leading to quality degradation in utterances tend to be linked with the spoof class, helping the model to classify such inputs as spoofed. The drop in **Te_B** performance may result from interventions causing bonafide utterances to resemble spoofed ones more closely. Second, the ratio always surpasses 1, as a universal trend regardless of which intervention applied. This emphasizes the bonafide class's modeling fragility, suggesting that future research should more focus on the robust modeling of the bonafide class.

4.3. Further test interventions with silence trimming

As introduced in Section 1, the silence could significantly influence the class-wise imbalance results. This prompts an important question: "Do our observations remain valid when silence is excluded from the training or testing datasets or both?" To address this, we conducted experiments with silence removed – referred to as silence trimming – to assess its effect on our findings and show the results in Table 4. The results show a remarkable consistency across most of our results. When silence was eliminated from both phases, effectively eliminating its influence, the ratio consistently exceeded 1, reaching a peak of 11.23. In additional experiments where silence was removed from only one phase, there were a few instances where the ratio fell below 1; however, the majority still exhibited ratios significantly greater than 1. These consistent outcomes robustly support the conclusion that the presence of silence does not skew our findings.

4.4. Asymmetric training results

Lastly, we focus on asymmetric interventions during the training phase to explore two key questions: (i) "What happens when interventions are applied solely during training?" and (ii) "Does the class-specific effect of interventions remain consistent if the test set remains untouched?"[2] According to the findings presented in Table 5, an improvement was observed in only one out of nine instances for the **Tr_S** condition. In contrast, interventions generally led to diminished performance in the bonafide class, highlighting the necessity for more advanced methods to accurately model genuine speech. Please note that we aim to compare the value between **Tr_B** and **Tr_S** for each condition to understand how each class is affected by each intervention, not the improvement from the baselines in the first two rows. Our findings diverge from those in [28], with EER deteriorating under all asymmetric intervention conditions except for one. This discrepancy is likely due to our one-side distinct strategy of applying interventions exclusively during either the training or testing phase, creating a mismatch. Nevertheless, similar to findings in [28], we also observed that interventions involving loudness were generally less impactful compared to other types of interventions across all evaluations.

5. Conclusion

This paper has conducted an in-depth investigation into the behavior of audio anti-spoofing models through various experiments focused on loss analysis and asymmetric interventions. Our analyses expand the perspective beyond attack-centric or silence-focused interpretations. The findings suggest that current training practices, which primarily aim to detect spoofing artifacts in known attacks, may neglect the robust modeling of bona fide speech, potentially introducing bias in model learning. By advocating for a more balanced focus on understanding both bona fide and spoofed classes, our research paves the way for future studies to enhance the efficacy of audio anti-spoofing systems.

[2]The second question is related to the analysis in Section 4.2, where we want to clarify whether performance degradation stems from suddenly appearing unseen domains in the test phase.

6. Acknowledgment

The work has been partially supported by the Academy of Finland (Decision No. 349605, project "SPEECHFAKES"). Experiments of this work used the Bridges2 system at PSC and Delta system at NCSA through allocations CIS210014 and IRI120008P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants #2138259,#2138286, #2138307, #2137603, #2138296.

7. References

[1] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015.

[2] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[3] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with Rawnet2," in *Proc. ICASSP*, 2021.

[4] 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," *Proc. Interspeech*, 2019.

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

[6] R. K. Das, J. Yang, and H. Li, "Data augmentation with signal companding for detection of logical access attacks," in *Proc. ICASSP*, 2021.

[7] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proc. ICASSP*, 2022.

[8] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *Proc. ICASSP*, 2020.

[9] D. Paul, M. Sahidullah, and G. Saha, "Generalization of spoofing countermeasures: A case study with asvspoof 2015 and btas 2016 corpora," in *Proc. ICASSP*, 2017.

[10] H.-j. Shim, J.-w. Jung, and T. Kinnunen, "Multi-dataset co-training with sharpness-aware optimization for audio anti-spoofing," in *Proc. Interspeech*, 2023.

[11] I. Himawan, F. Villavicencio, S. Sridharan, and C. Fookes, "Deep domain adaptation for anti-spoofing in speaker verification systems," *Computer Speech & Language*, vol. 58, pp. 377–402, 2019.

[12] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Transactions on Information Forensics and Security*, 2023.

[13] X. Wang and J. Yamagishi, "Investigating active-learning-based training data selection for speech spoofing countermeasure," in *Proc. SLT*, 2022.

[14] 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," in *Proc. Interspeech*, 2022.

[15] X. Wang and J. Yamagishi, "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?" in *Proc. ICASSP*, 2024.

[16] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," *Proc. Speaker Odyssey Workshop*, 2020.

[17] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE SPL*, vol. 28, pp. 937–941, 2021.

[18] Y. Zhang, Z. Li, J. Lu, H. Hua, W. Wang, and P. Zhang, "The impact of silence on speech anti-spoofing," *IEEE/ACM-T-ASLP*, 2023.

[19] Y. Ren, H. Peng, L. Li, and Y. Yang, "Lightweight voice spoofing detection using improved one-class learning and knowledge distillation," *IEEE Transactions on Multimedia*, 2023.

[20] B. Chettri, R. G. Hautamäki, M. Sahidullah, and T. Kinnunen, "Data quality as predictor of voice anti-spoofing generalization," in *Proc. Interspeech*, 2021.

[21] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM-T-ASLP*, 2023.

[22] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[23] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017.

[24] B. Chettri, E. Benetos, and B. L. Sturm, "Dataset artefacts in anti-spoofing systems: a case study on the asvspoof 2017 benchmark," *IEEE/ACM-T-ASLP*, vol. 28, pp. 3018–3028, 2020.

[25] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019.

[26] N. M. Müller *et al.*, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?" *Proc. ASVspoof 2021 Workshop (Interspeech satellite)*, pp. 55–60, 2021.

[27] Y. Zhang, W. Wang, and P. Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, 2021.

[28] H.-j. Shim, R. G. Hautamäki, M. Sahidullah, and T. Kinnunen, "How to construct perfect and worse-than-coin-flip spoofing countermeasures: A word of warning on shortcut learning," in *Proc. Interspeech*, 2023.

[29] L. Wu, Z. Zhu *et al.*, "Towards understanding generalization of deep learning: Perspective of loss landscapes," *ICML Workshop*, 2017.

[30] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin, "Learning from failure: De-biasing classifier from biased classifier," *in Proc. NeurIPS*, vol. 33, pp. 20 673–20 684, 2020.

[31] J. Lee, E. Kim, J. Lee, J. Lee, and J. Choo, "Learning debiased representation via disentangled feature augmentation," *in Proc. NeurIPS*, vol. 34, pp. 25 123–25 133, 2021.

[32] I. Hwang, S. Lee, Y. Kwak, S. J. Oh, D. Teney, J.-H. Kim, and B.-T. Zhang, "Selecmix: Debiased learning by contradicting-pair sampling," *in Proc. NeurIPS*, vol. 35, pp. 14 345–14 357, 2022.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[34] T. Castells, P. Weinzaepfel, and J. Revaud, "Superloss: A generic loss for robust curriculum learning," *in Proc. NeurIPS*, vol. 33, pp. 4308–4319, 2020.

[35] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *Proc. CVPR*, 2020.

[36] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *in Proc. NeurIPS*, vol. 31, 2018.

[37] J. Yamagishi, X. Wang *et al.*, "ASVspoof2021: Accelerating progress in spoofed and deep fake speech detection," in *Proc. ASVspoof 2021 Workshop (Interspeech satellite)*, 2021.

[38] 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.

[39] A. Y. Kibangou and G. Favier, "Wiener-Hammerstein systems modeling using diagonal volterra kernels coefficients," *IEEE SPL*, vol. 13, no. 6, pp. 381–384, 2006.

[40] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *in Proc. ICLR*, 2020.

[41] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *Proc. ICML*, 2021.

[42] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[43] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.