ENHANCING MULTI-CORPUS TRAINING IN SSL-BASED ANTI-SPOOFING MODELS: DOMAIN-INVARIANT FEATURE EXTRACTION

Anh-Tuan DAO, Driss Matrouf, Mickael Rouvier

Laboratoire d'informatique d'Avignon, France

Nicholas Evans

EURECOM, Sophia Antipolis, France

ABSTRACT

The performance of speech spoofing detection often varies across different training and evaluation corpora. Leveraging multiple corpora typically enhances robustness and performance in fields like speaker recognition and speech recognition. However, our spoofing detection experiments show that multi-corpus training does not consistently improve performance and may even degrade it. We hypothesize that dataset-specific biases impair generalization, leading to performance instability. To address this, we propose an Invariant Domain Feature Extraction (IDFE) framework, employing multi-task learning and a gradient reversal layer to minimize corpus-specific information in learned embeddings. The IDFE framework reduces the average equal error rate by 20% compared to the baseline, assessed across four varied datasets.

Index Terms— deepfake detection, domain-invariant

1. INTRODUCTION

Automatic Speaker Verification (ASV) offers reliable biometric recognition based on voice characteristics. However, advanced spoofing attacks that mimic bona fide users pose serious security risks to ASV systems. To counter these threats, spoofing detection systems, also commonly referred to as countermeasures (CMs), are increasingly deployed to protect ASV reliability. In the biometrics literature, this task is formally described as Presentation Attack Detection (PAD), where the goal is to discriminate between genuine presentations and attack presentations.[1] The ASVspoof [1] and ADD [2] communities, with broad global participation, have made substantial advances in this field.

Recent research has focused on leveraging self-supervised learning (SSL) models to enhance spoofing detection. Fine-tuning SSL representations has become the dominant paradigm, as their high-dimensional acoustic features provide superior generalization over traditional supervised end-to-end architectures [3, 4, 5]. For instance, Tak et al. [6] fine-tuned the wav2vec 2.0 model, a large-scale SSL-based speech representation framework, using the ASVspoof 2019 LA dataset. Their approach demonstrated improved robustness against unseen spoofing attacks when evaluated using the ASVspoof 2021 LA and DF datasets. Roselló et al. [7] utilized Conformer architectures to capture both local and global speech patterns, achieving state-of-the-art performance using the same datasets. The Multi-Head Factorized Attentive (MHFA) pooling mechanism, originally developed for SSL-based speaker verification [8], has been adapted into a lightweight architecture for aggregating transformer embeddings achieving superior performance in the ASVspoof 5 Challenge [9].

Despite these advancements, spoofing detection systems exhibit inconsistent performance due to dataset-specific biases in training data, as noted by Nicolas et al. [10], who highlighted the impact of non-speech segments on model predictions. To address these challenges, we conducted experiments to improve model reliability in the future of dataset variability. We utilized the ASVspoof 2019 [11], ASVspoof 5 [1], and Fake-or-Real (FoR) [12] datasets to create diverse training scenarios.

In fields such as speaker recognition and speech recognition, using multiple corpora is generally known to enhance robustness and improve performance. In contrast, our experiments in spoofing detection show that the use of multiple corpora does not consistently improve performance; in some cases, it significantly degrades performance. Analysis of the embedding space learned under multi-corpus conditions reveals distinct clustering patterns aligned with dataset IDs, indicating the capture of corpus-specific information, potentially leading to overfitting and reduced generalization on unseen data due to dataset-specific biases. We hypothesize that these biases hinder generalization and contribute to performance instability.

Mitigating these biases could improve generalization, enabling models to leverage multiple datasets. While the Gradient Reversal Layer (GRL) is a proven mechanism for mitigating domain shifts in general applications [13], it has also demonstrated efficacy in spoofing detection for suppressing non-discriminative attributes, such as speaker-specific identities [14]. We propose the *Invariant Domain Feature Extraction* (IDFE) framework, which incorporates domain-adversarial training with a GRL to suppress dataset-specific cues in the embedding space. This approach enhances the focus on discriminative spoofing-related features leading to

[1] In this paper, the term *spoofing detection* is used interchangeably with *presentation attack detection*. Similarly, *bona fide* audio corresponds to *genuine presentations*, while *spoofing audio* refer to *attack presentations*.

consistent improvement in detection performance.

Our main contributions are:

- **Multi-Corpus Training and Dataset Bias Analysis**: We use four training scenarios including multi-corpus training datasets and four evaluation datasets to assess reliability. Our findings highlight significant performance variability due to dataset bias, revealing limitations in the generalization of current spoofing detection systems.
- **Invariant Domain Feature Extraction (IDFE)**: We propose the IDFE framework that employs domain-adversarial training to remove corpus information from embeddings, improving generalization to unseen data distributions and reducing the impact of dataset bias on model performance. The IDFE framework achieves a 20% reduction in average EER over four evaluation datasets compared to the MHFA baseline and can be applied to other detection architectures.

2. SSL-BASED SPOOFING DETECTION

Our spoofing detection approach uses SSL models to extract robust audio features. It includes two components: SSL-based feature extractors and back-end classifiers.

2.1. SSL Feature Extractors

SSL models learn meaningful representations from unlabeled data through pretext tasks, capturing intrinsic patterns. In speech processing, SSL models like wav2vec 2.0 excel at extracting high-quality features from raw audio. The model consists of a convolutional neural network (CNN) feature encoder and a Transformer-based context network. The feature encoder processes raw audio into latent representations $z_{1:T}$, reducing dimensionality and capturing acoustic features. These representations are then input into a Transformer-based context network, which models temporal dependencies and contextual information, producing context representations $o_{1:T}$. The Transformer's self-attention mechanism enables the model to capture long-range relationships, essential for understanding speech structure.

2.2. Back-end Classifiers

We use three back-end classifiers as baselines:

- **AASIST** [6]: is specifically designed for spoofing detection, integrating spectro-temporal feature extraction with deep learning. It leverages convolutional and attention-based mechanisms to capture discriminative patterns in audio signals.
- **Conformer** [7]: combines convolutional and Transformer layers, effectively modeling both local and global dependencies in speech signals. This hybrid structure has shown strong performance in spoofing detection.
- **MHFA** [9]: Multi-Head Factorized Attentive (MHFA) pooling is an attention-based backend designed to efficiently aggregate Transformer embeddings using multi-head attention mechanisms with factorized attentive pooling. Originally developed for SSL-based speaker verification, MHFA employs a lightweight yet effective architecture to enhance feature representation.

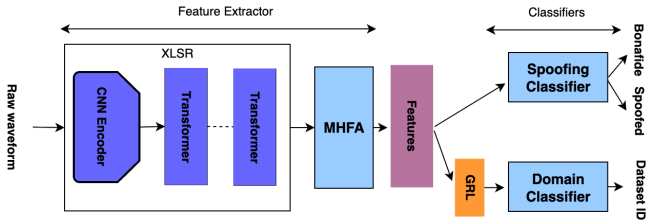


Fig. 1: Our model architecture with dual classifiers and GRL.

3. INVARIANT DOMAIN FEATURE EXTRACTION

To mitigate the negative impact of training corpus-specific artifacts, we introduce the IDFE framework. Our approach encourages the model to discard dataset-dependent information from its learned embeddings, encouraging it to focus on attack-related, discriminative cues.

3.1. Architecture

Our IDFE framework builds upon the SSL-based backbone and MHFA module. As illustrated in Figure 1, the model consists of the following components.

- **Feature Extractor:** A pre-trained XLSR [15] SSL encoder to extract contextualized frame-level embeddings from raw audio, combined with a MHFA module to aggregate contextualized embeddings using attention mechanisms.
- **Spoofing Classifier Head:** A fully connected network comprising linear layers with batch normalization, ReLU activations, and dropout, which use MHFA-aggregated embeddings to predict the binary label $\hat{y}_s \in \{\text{bona fide}, \text{spoof}\}$.
- **Domain Classifier Head:** A parallel fully connected network (structurally similar to the spoofing head), which predicts the dataset ID $\hat{y}_d \in 1, \ldots, D$, where D is the number of training corpora.

- **Gradient Reversal Layer (GRL):** Inserted between the feature extractor and the domain classifier head. In the forward pass, the GRL acts as an identity. During backpropagation, it multiplies the gradients by a hyper parameter $-\lambda$, thus encouraging the feature extractor to remove domain-specific patterns.

3.2. Domain Adversarial Training

To induce domain-invariant embeddings, we adopt a domain adversarial objective. Let $o = f(x)$ denote the feature sequence for the input audio x, and let $g_s(\cdot)$ and $g_d(\cdot)$ be the spoofing and domain classifiers, respectively. During training, we optimize:

$$\min_{f,\, g_s\, g_d} \mathcal{L}_s\big(g_s(f(x)),\, y_s\big) \;+\; \alpha\, \mathcal{L}_d\big(g_d\big(\mathrm{GRL}(f(x))\big),\, y_d\big) \tag{1}$$

where, $\mathcal{L}_s$ is the cross-entropy loss for spoofing classification. $\mathcal{L}_d$ is the cross-entropy loss for domain classification (dataset ID). The hyper parameter α is used to balance the two objectives, controlling the strength of domain invariance.

3.3. Backpropagation and Parameter Updates

During training, the model parameters are updated using stochastic gradient descent. Let θ_f, θ_{g_s}, and θ_{g_d} represent the parameters of the feature extractor, spoofing classifier, and domain classifier, respectively. The parameter updates are performed as follows:

$$\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial \mathcal{L}_s}{\partial \theta_f} - \alpha\lambda \frac{\partial \mathcal{L}_d}{\partial \theta_f} \right) \tag{2}$$

$$\theta_{g_s} \leftarrow \theta_{g_s} - \mu \frac{\partial \mathcal{L}_s}{\partial \theta_{g_s}} \tag{3}$$

$$\theta_{g_d} \leftarrow \theta_{g_d} - \mu \frac{\partial \mathcal{L}_d}{\partial \theta_{g_d}} \tag{4}$$

In Equations (2-4), μ denotes the learning rate, whereas λ is a trade-off hyperparameter in GRL which controls the influence of the domain loss [13]. Equation 2 updates the feature extractor to reduce the spoofing loss $\mathcal{L}_s^i$ while simultaneously increasing the domain loss $\mathcal{L}_d^i$, thereby confusing the domain classifier while encouraging the spoofing classifier to learn spoofing-related cues. This is achieved through the GRL, which multiplies the gradient from the domain classifier by $-\lambda$ during backpropagation. Equations 3 and 4 update the spoofing and domain classifiers, respectively, using standard gradient descent to minimize $g_s(\cdot)$ and $g_d(\cdot)$.

4. EXPERIMENTAL SETUP

4.1. Datasets and Metrics

For training, we use the ASVspoof 5 training set, the ASVspoof 2019 [11] and FoR [12] datasets. To support multi-corpus evaluation, the ASVspoof 5 validation set is replaced with the ITW dataset. We merge the ASVspoof 2019 training and development sets to increase data volume, ensuring evaluation integrity, our evaluation sets remain disjoint and unaffected by development data leakage. We define four distinct training scenarios:

- Case 1: ASVspoof 5 training set only.
- Case 2: ASVspoof 2019 train-dev set.
- Case 3: Combination of ASVspoof 5 and ASVspoof 2019.
- Case 4: Combination of ASVspoof 5, ASVspoof 2019, FoR.

For evaluation, we use three datasets: ASVspoof 2021 LA and DF hidden subsets [16, 17] and ASVspoof 5 evaluation set. The ASVspoof 2021 LA and DF hidden subsets have non-speech segments removed, preventing models from exploiting such shortcuts. Prior work [18] shows the challenges presented by these hidden subsets, compared to standard evaluation subsets, since they demand a reliance on core spoofing cues rather than non-speech-related shortcuts.

Performance is evaluated using the equal error rate (EER). In biometric evaluation practice, system performance is often reported in terms of the attack and bona fide presentation classification error rates (APCER and BPCER). In this paper, the EER corresponds to the operating point at which these two error rates are equal.

4.2. Data Augmentation

We applied standard data augmentation techniques during training, utilizing the MUSAN corpus and the real room impulse response (RIR) database [19, 20]. Each training utterance underwent one of four augmentation methods:

- Reverberation: Utterances are convolved with real RIRs to simulate reverberation effects associated with propagation in various acoustic spaces.
- Speech: A summation of three to eight different-speaker utterances are added to each training utterance at signal-to-noise ratios (SNRs) of 13-20 dB.
- Music: Randomly-selected music recordings from MUSAN are added to each training utterance at SNRs of 5-15 dB.

Table 1: Model performance of baselines across four training cases. The *pooled* indicates the average EER across all evaluation datasets. Details of the datasets used in each training case are provided in Section 4.1.

		EER				
Training case	**Model**	**ITW**	**ASV5 eval**	**ASV21LA-Hidden**	**ASV21DF-Hidden**	**Pooled**
Case 1	AASIST	7.03	5.54	13.66	9.6	8.95
	Conformer	5.69	3.85	12.49	10.4	8.10
	MHFA	4.31	4.64	12.14	8.58	7.41
Case 2	AASIST	4.05	12.35	8.9	8.35	8.41
	Conformer	3.36	15.12	7.17	7.27	8.23
	MHFA	2.67	12.68	9.33	7.27	7.98
Case 3	MHFA	2.27	8.72	7.57	5.91	6.11
	MHFA-IDFE	**2.04**	**7.97**	**5.22**	**4.29**	**4.88**
Case 4	MHFA	2.33	8.36	6.80	4.44	5.48
	MHFA-IDFE	**1.86**	**7.95**	**5.40**	**3.58**	**4.69**

Table 2: Performance for MHFA-IDFE with different values of α which controls the strength of domain invariance.

Training case	**Model**	α	**ITW**	**ASV5 eval**	**ASV21LA-Hidden**	**ASV21DF-Hidden**	**Pooled**
	MHFA-IDFE	0.1	1.86	7.95	5.40	3.58	4.69
Case 4	MHFA-IDFE	0.5	2.31	9.10	6.28	4.04	5.43

- Noise: Randomly-selected noise recordings from MUSAN are added to each training utterance at SNRs of 0-15 dB.

4.3. Implementation Details

To mitigate the non-speech bias in model training, we trim non-speech segments from the beginning and end of training utterances using the `librosa` library. Specifically, we apply `librosa.effects.trim` with a reference threshold of 40 dB to determine the duration of leading and trailing non-speech segments. Training utterances are randomly segmented into 4-second clips. For evaluation, models operate on full utterances. We use the Adam optimizer [21] with a learning rate of 10^{-6} to minimize the weighted cross-entropy loss, adjusted based on the ratio of bona fide to spoofed audio samples. The hyperparameter α is set to 0.1 to control the strength of domain invariance. The hyperparameter λ gradually changes from 0 to 1, following the same schedule as in [13]. Training is conducted in batches of 32 samples over 30 epochs using NVIDIA A100 GPUs.

5. RESULTS

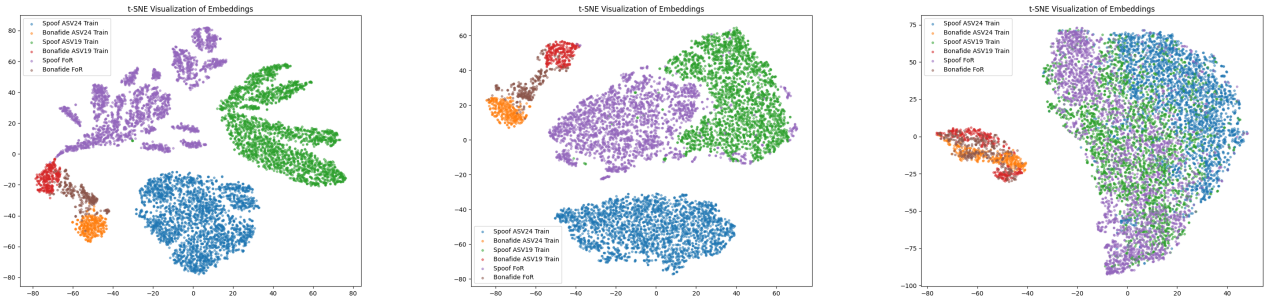
We present an evaluation of three baseline models across two training scenarios, highlighting variability due to training and evaluation datasets. Subsequently, we select the best-performing baseline, MHFA, to integrate with our proposed IDFE framework for multi-corpus training.

5.1. Dataset Bias Analysis

A performance comparison for models trained using ASVspoof 5 and ASVspoof 2019 presented in Table 1 (Case 1-2) shows significant variations that highlight dataset bias issues. For instance, the MHFA model trained using ASVspoof 5 (Case 1) achieves an EER of 4.64% for the ASVspoof 5 evaluation dataset while performance degrades to 12.14% for the ASVspoof 2021 LA dataset. Training using ASVspoof 2019 train-dev (Case 2) yields an EER of 9.33% for the ASVspoof 2021 LA dataset and 12.68% for the ASVspoof 5 dataset. Similar trends are observed for AASIST and Conformer models, underscoring vulnerability to out-of-domain data. For multi-corpus training (Case 3), the combination of ASVspoof 5 and ASVspoof 2019, the MHFA model shows improved generalization for all datasets (e.g., ITW, ASVspoof 2021 LA, and DF). However, performance for the ASVspoof 5 dataset drops from 4.64% to 8.72% EER, indicating sensitivity to dataset composition. Among the baselines, MHFA demonstrates superior performance, making it the candidate for applying our IDFE framework in multi-corpus training.

5.2. Multi-Corpus Training Performance

We evaluate the MHFA-IDFE model, incorporating our IDFE framework into the MHFA baseline in two multi-corpus train-



(a) MHFA. (b) MHFA-IDFE (α = 0.1). (c) MHFA-IDFE (α = 0.5).

Fig. 2: t-SNE visualization of learned embeddings from the MHFA and MHFA-IDFE models for the training case 4. The parameter α controls the strength of domain invariance, as defined in Equation 1.

ing scenarios (Case 3-4 in Table 1). In Case 3 (ASVspoof 5 + ASVspoof 2019), MHFA-IDFE performance surpasses that of MHFA across all evaluation datasets, reducing the pooled EER by 20% (from 6.11% to 4.88%). The largest improvement is for the ASVspoof 2021 LA, with a 31% EER reduction (from 7.57% to 5.22%). In Case 4 (ASVspoof 5, ASVspoof 2019, and FoR), MHFA-IDFE achieves a 14.4% performance gain, reducing the pooled EER from 5.48% to 4.69%. Notable improvements occur for ITW and ASVspoof 2021 LA datasets, with EER reductions of 20.1% and 20.5%, respectively. Although MHFA-IDFE improves performance for ASVspoof 5 in Case 4 (EER from 8.36% to 7.95%), the EER remains higher than that in Case 1. This indicate that while MHFA-IDFE mitigates the dataset bias issue, it does not fully resolve it.

5.3. Visualization

A visualisation of MHFA and MHFA-IDFE embeddings for Case 4 derived using t-SNE is inlustrated in (Fig 2). For the MHFA model, there is a distinct clustering by dataset, suggesting the model captures corpus-specific cues. With $\alpha = 0.1$, MHFA-IDFE yields embeddings with compact dataset-based clusters and clear separation between bona fide and spoofed samples; spoofed samples from FoR and ASVspoof 2019 are closer but still separable. The MHFA-IDFE with $\alpha = 0.5$ produces embeddings with no evident dataset clustering, though performance is inferior to $\alpha = 0.1$ (Table 2), possibly due to excessive influence from the invariant domain loss. This excessive influence can suppress task-relevant information, such as attack type, reducing the model ability to distinguish bona fide from spoofed samples. Balancing the strength of invariant domain loss is therefore crucial to retain discriminative features while mitigating dataset bias.

6. CONCLUSION

By investigating spoofing detection in multi-corpus training, this study shows dataset bias issues, notably in the ASVspoof 5 evaluation set, emphasizing the necessity for bias mitigation. Our proposed IDFE framework fosters domain-invariant embeddings, achieving a 20% reduction in pooled EER across four evaluation datasets. The framework effectively removes corpus-specific information from the learned embedding space. Balancing the impact of the invariant domain loss is essential to prevent excessive information loss, which could impair spoofing detection performance. Notably, this work emphasizes domain robustness rather than optimizing peak performance on a single backend architecture; consequently, direct comparisons with results of prior works, often trained and evaluated under different protocols, are infeasible.

7. ACKNOWLEDGEMENTS

This work was performed using HPC resources from GENCI-IDRIS. This work was financially supported by ANR BRUEL (ANR-22-CE39-0009).

8. REFERENCES

[1] Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi, "ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *ASVspoof 2024*, 2024.

[2] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang,

Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li, "ADD 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022*, 2022.

[3] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022*. 2022, pp. 6367–6371, IEEE.

[4] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021*. 2021, pp. 6369–6373, IEEE.

[5] Anh-Tuan Dao, Mickael Rouvier, and Driss Matrouf, "ASVspoof 5 challenge: advanced resnet architectures for robust voice spoofing detection," in *ASVspoof 2024*, 2024, pp. 163–169.

[6] Tak Hemlata, Todisco Massimiliano, Wang Xin, Jung Jee-weon, Yamagishi Junichi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Odyssey 2022*, 2022.

[7] Eros Rosello, Alejandro Gómez Alanís, Angel M. Gomez, and Antonio M. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Interspeech 2023*, 2023.

[8] Junyi Peng, Oldrich Plchot, Themos Stafylakis, Ladislav Mosner, Lukás Burget, and Jan Cernocký, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *IEEE Spoken Language Technology Workshop, SLT 2022*, 2022.

[9] Theophile Stourbe, Victor Miara, Théo Lepage, and Réda Dehak, "Exploring WavLM back-ends for speech spoofing and deepfake detection," in *ASVspoof 2024*, 2024.

[10] Nicolas Michael Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, and Konstantin Böttinger, "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?," *ArXiv*, 2021.

[11] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, 2019.

[12] Ricardo Reimao and Vassilios Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019, Timisoara, Romania, October 10-12, 2019*, Corneliu Burileanu and Horia-Nicolai Teodorescu, Eds., 2019.

[13] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[14] Anh-Tuan Dao, Driss Matrouf, and Nicholas Evans, "Assessing the impact of speaker identity in speech spoofing detection," 2026.

[15] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022.

[16] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2023.

[17] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[18] Juan M. Martín-Doñas, Aitor Álvarez, Eros Rosello, Angel M. Gomez, and Antonio M. Peinado, "Exploring Self-supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection," in *Interspeech 2024*, 2024.

[19] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *ArXiv*, 2015.

[20] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP 2017*, 2017.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, Yoshua Bengio and Yann LeCun, Eds., 2015.