Odyssey 2026
23-26 June 2026, Lisbon, Portugal

A Comparison of SSL-Based Feature Extractors and Back-End Classifiers for Spoofing Detection: A Multi-Corpus Training and Cross-Linguistic Analysis

Anh-Tuan Dao[1], *Driss Matrouf*[1], *Mickael Rouvier*[1], *Nicholas Evans*[2]

[1]Laboratoire Informatique d'Avignon, Avignon Universite, France
[2]EURECOM , Sophia Antipolis, France

{anh-tuan.dao, driss.matrouf, mickael.rouvier}@univ-avignon.fr, evans@eurecom.fr

Abstract

Voice biometric systems face growing threats from spoofing attacks, yet the evaluation of detection models remains inconsistent across datasets. To investigate these unpredictable fluctuations, we conduct a comprehensive benchmark of four self-supervised learning feature extractors paired with four back-end classifiers. We compare the hierarchical local feature extraction of ResNet with the global sequence and relational modeling of attention and graph-based back-ends. Through multi-corpus training across three scenarios and six evaluation datasets, our empirical analysis yields two critical findings. First, we expose a domain bias within the ASVspoof 5 dataset, showing that naive data scaling actively degrades performance. Second, our cross-linguistic analysis reveals that fine-tuning with just 8 hours of target-language data enhances detection robustness. Together, these findings emphasize the critical need for domain-aware and language-specific adaptation in spoofing detection.

1. Introduction

Automatic Speaker Verification (ASV) offers a reliable and convenient biometric recognition method based on voice characteristics. However, advanced spoofing attacks which mimic bona fide users pose serious security risks. To counter these threats, spoofing detection systems, also called countermeasures (CMs), are increasingly deployed to protect ASV integrity. The ASVspoof [1] community, with global participation, has driven major advancements in this field.

Recent years have witnessed a paradigm shift in spoofing detection with the adoption of self-supervised learning (SSL) models [2, 3, 4, 5, 6]. By exploiting large-scale pre-training with raw speech, SSL-based models learn expressive representations which outperform conventional supervised architectures [7, 8, 9]. Despite these gains, robust generalization across datasets and attack types remains a challenge. Models trained using one particular corpus often exhibit sharp performance degradations when evaluation is performed with other corpora.

A conventional hypothesis in machine learning is that incorporating additional, diverse data into the training pipeline should yield a more generalized and robust model. However, in the context of spoofing detection, our investigations reveal a counter-intuitive phenomenon: expanding the training set with additional spoofing corpora frequently leads to performance degradation on unseen evaluation sets. This observation suggests the presence of hidden dataset biases that override generalized spoofing cues. To diagnose the underlying mechanics of this performance degradation, we visualized the learned representations using t-SNE. The projections reveal a critical vulnerability in multi-corpus training: alongside the expected separation of bona fide and spoofed utterances, the embeddings exhibit distinct clustering driven by their source dataset. This strong domain separation provides empirical evidence of dataset bias, demonstrating that the network could overfit to dataset-specific shortcuts rather than extracting generalizable spoofing artifacts.

Despite advances in spoofing back-end classifiers, the integration of deep convolutional architectures, specifically ResNets, with SSL-based feature extractors remains largely underexplored. Recent literature heavily favors attention-based sequence aggregators (such as MHFA and Conformer) or specialized graph-based models (like AASIST). While highly effective, attention mechanisms primarily capture global sequence dependencies, whereas graph-based models focus on relational structures. In contrast, ResNets provide a hierarchical extraction of local features. Given the complex, high-dimensional nature of the representations generated by modern SSL front-ends, we hypothesize that processing these embeddings through a robust convolutional architecture could capture localized spoofing artifacts that global aggregators might miss. To investigate this, we systematically compare ResNet against AASIST, Conformer, and MHFA. To this end, we conduct a comprehensive evaluation of multiple SSL feature extractors paired with diverse back-end classifiers.

Our primary contributions are as follows:

- **Empirical Diagnosis of Dataset Bias in Multi-Corpus Training -** Using t-SNE visualizations, we provide empirical evidence that naive multi-corpus training could causes models to rely on dataset-specific shortcuts, clustering embeddings by source domain.
- **Performance Comparison of Back-end Classifiers -** We compare the performance of the ResNet back-end against three back-ends, including AASIST, Conformer, and MHFA. This comparative analysis benchmarks the empirical effectiveness of a convolutional, local-feature approach against global sequence aggregators and graph-based models when processing high-dimensional SSL representations.
- **Performance Comparison of Front-end Classifiers -** We isolate the impact of the front-end by evaluating four SSL feature extractors (Wav2vec2, HuBERT, WavLM, and XLSR). Our findings show that the scale and diversity of pre-training data are directly correlated with generalization, multilingual pre-training (XLSR) yielding significantly more robust representations for spoofing detection.
- **Cross-linguistic Evaluation -** We extend our investigation to cross-linguistic scenarios by evaluating spoofing detection performance using Spanish (HABLA dataset) and Chinese (CFAD dataset) language data. Our findings indicate that even a modest inclusion of target language training data leads to marked improvements in detection performance, underscoring the importance of language-specific adaptation.

10.21437/Odyssey.2026-39

2. SSL-based Spoofing Detection

Our approach to spoofing detection leverages the strengths of SSL models to extract robust representations from raw audio data. Models are comprised of two main components: SSL-based feature extractors and back-end classifiers (Figure 1).

2.1. SSL Feature Extractors

The model comprises two primary components: a convolutional neural network (CNN) feature encoder and a Transformer-based context network. The feature encoder role is to process the raw audio input and transform it into a sequence of latent representations $z_{1:T}$. This transformation reduces the dimensionality of the input data and captures acoustic features. Following the feature encoder, the sequence of latent representations $z_{1:T}$, is input into a Transformer-based context network. The purpose of this network is to model the temporal dependencies and contextual information within the audio sequence, producing context representations $o_{1:T}$. The Transformer architecture, known for its self-attention mechanism, allows the model to capture long-range relationships in the data, which is crucial for understanding the structure of speech. In this study, we investigate four widely used SSL models for spoofing detection:

- **Wav2Vec2** [10]: consists of a convolutional feature encoder and a Transformer-based context network. The feature encoder converts raw speech into latent representations, while the Transformer refines these embeddings by capturing long-range dependencies, producing contextualized representations.
- **HuBERT** [11]: differs from Wav2Vec2 in its training objective. Instead of contrastive learning, it employs masked prediction with discrete target representations derived from clustering.
- **WavLM** [12]: extends Wav2Vec2 by incorporating gated relative position bias and a denoising training objective, making it more robust to noise and distortions.
- **XLSR** [13]: is a multilingual variant of Wav2Vec2, pre-trained on huge speech data from multiple languages.

Each SSL model captures speech features differently due to variations in architecture, training objectives, and pre-training datasets. By evaluating and comparing these models, we aim to identify which SSL-based features are most effective for spoofing detection.

2.2. Back-end Classifiers

We evaluate and compare three back-end classifiers and our proposed ResNet back-end classifier:

- **AASIST** [14]: 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** [3]: 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** [4]: 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.
- **ResNet**: We propose a ResNet-based back-end classifier which takes as input the final embedding produced by the SSL model. Our ResNet architecture consists of 34 layers with residual connections. The model utilizes a basic block architecture characterized by two 3×3 convolutional layers, with each convolution followed by a batch normalization layer and ReLU activation. Each basic block is followed by a dropout layer with a rate of 0.5 to prevent overfitting. The detailed architecture of our ResNet model is summarized in Table 1.

By evaluating these back-end classifiers, we aim to identify the most effective architecture for integrating SSL-based representations in spoofing detection tasks.

Layer name	Output shape	Input
Data-aug	64600	
SSL front-end	1024 x T	Wav2vec 2.0
Conv2D	32 x 1024 x T	$[3 \times 3, 32]$
BasicBlock-1	32 x 512 x T	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$
BasicBlock-2	64 x 256 x T/2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$
BasicBlock-3	128 x 128 x T/4	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$
BasicBlock-4	256 x 64 x T/8	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$
Flatten	16384 x T/8	
Mean Pooling	16384	
Dense	2	

Table 1: *The architecture of our ResNet back-end classifier with SSL feature extractor.*

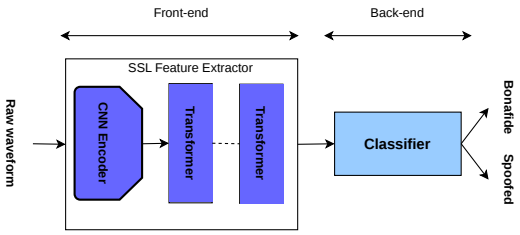


Figure 1: *SSL-based spoofing detection model architecture.*

3. Experimental Setup

3.1. Datasets and Metrics

For training, we use the ASVspoof 5 training set, MLAAD-v3 [16], ASVspoof19 [17], and VCTK [18]. To support multi-corpus evaluation, the ASVspoof 5 validation set is replaced

Table 2: *Performance of back-end classifiers with the XLSR feature extractor in three training cases (best results highlighted in bold, Average EER across five evaluation datasets in last column). Hidden subsets of ASVspoof21 LA and DF [15] are used for evaluation.*

Training	Back-end	Wild		ASV5 eval		ASV21 LA Hidden		ASV21 DF Hidden		Fake-Or-Real		Average
		EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER
Case 1	AASIST	8.72	0.20	6.41	0.19	11.48	0.26	8.43	0.20	5.91	0.14	8.19
	Conformer	5.23	0.14	5.19	0.15	12.09	0.29	9.24	0.23	**2.10**	**0.06**	6.77
	MHFA	4.71	0.13	5.56	0.16	**10.86**	**0.25**	8.63	0.20	6.66	0.17	7.28
	ResNet	**3.96**	**0.11**	**4.73**	**0.14**	11.28	0.25	**8.24**	**0.18**	4.38	0.12	**6.52**
Case 2	AASIST	2.72	0.07	**10.24**	**0.30**	10.71	0.28	8.39	0.22	0.59	0.02	6.53
	Conformer	4.02	0.12	12.49	0.35	11.07	0.26	8.08	0.20	0.17	0.00	7.16
	MHFA	2.40	0.06	11.04	0.32	9.60	0.23	6.71	0.17	0.26	0.01	6.00
	ResNet	**2.02**	**0.05**	12.37	0.36	**8.21**	**0.21**	**5.97**	**0.16**	**0.17**	**0.00**	**5.75**
Case 3	AASIST	1.80	0.05	13.69	0.39	5.72	0.16	3.93	0.10	0.48	0.01	5.12
	Conformer	1.69	0.04	13.35	0.39	**4.59**	0.11	**2.57**	**0.06**	0.26	0.01	4.49
	MHFA	1.45	0.03	12.06	0.35	5.86	0.13	3.53	0.07	0.26	0.01	4.63
	ResNet	**1.21**	**0.03**	**11.46**	**0.32**	5.21	**0.11**	3.03	0.07	**0.13**	**0.00**	**4.20**

Table 3: *Performance Comparison of SSL feature extractors with our ResNet back-end in training case 2.*

Extractor	Wild		ASV5 eval		ASV21 LA Hidden		ASV21 DF Hidden		Fake-Or-Real		Average
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER
Wav2vec2	9.03	0.26	**7.41**	**0.21**	19.49	0.50	15.52	0.41	1.67	0.05	10.62
HuBERT	15.13	0.42	9.36	0.27	33.47	0.74	29.39	0.64	3.54	0.09	18.18
WavLM	7.82	0.17	10.12	0.29	14.74	0.36	10.26	0.25	0.39	0.01	8.67
XLSR	**2.02**	**0.05**	12.37	0.36	**8.21**	**0.21**	**5.97**	**0.16**	**0.17**	**0.00**	**5.75**

with the ITW dataset. We define three distinct training scenarios:

- Training Case 1: ASVspoof 5 training set only.
- Training Case 2: ASVspoof 5 training and MLAAD-v3.
- Training Case 3: ASVspoof 5 training set, MLAAD-v3, ASVspoof19, and VCTK.

For evaluation, we use six evaluation datasets: four English (ASVspoof21 LA and DF Hidden [15, 19], ASVspoof 5 evaluation set, Fake-Or-Real [20]) and two non-English (HABLA [21], CFAD [22] noisy-unseen-test). The ASVspoof 2021 LA and DF hidden subsets have non-speech segments removed, preventing models from exploiting such shortcuts. Prior work [23] 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 measured using Equal Error Rate (EER) and minimum Detection Cost Function (minDCF).

3.2. Data Augmentation

We apply standard data augmentation techniques during training, utilizing the MUSAN corpus and the real room impulse response (RIR) database [24, 25]. Each training utterance underwent 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.
- Noise: Randomly-selected noise recordings from MUSAN are added to each training utterance at SNRs of 0-15 dB.

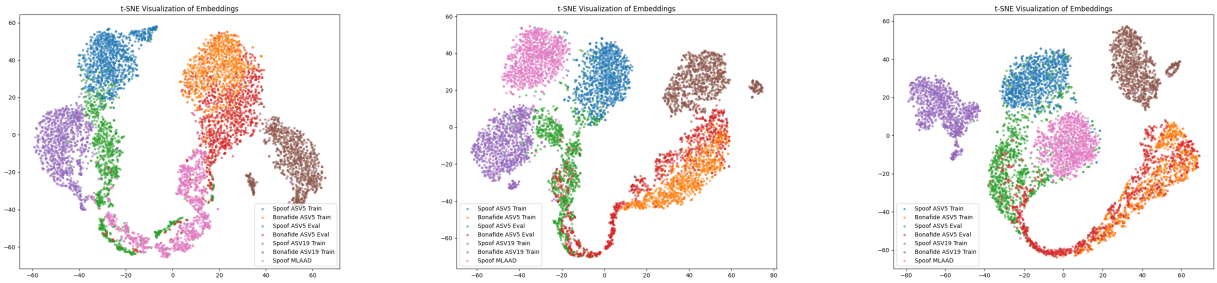
3.3. Implementation Details

All models, including the SSL front-ends and back-end classifiers, are fine-tuned end-to-end. During training, utterances are randomly segmented into 4-second clips before being fed into the models, whereas evaluation is performed on the complete audio clips. We employ the standard Adam [26] optimizer with a learning rate of 10^{-6}, a weight decay of 10^{-5}, in order to minimize a weighted cross-entropy loss function. Training is performed in batches of 32 samples over 30 epochs using Nvidia A100 GPUs, with the best checkpoint selected.

We evaluate four SSL feature extractors from the Torchaudio library: Wav2Vec2-Large-LV60K, HuBERT-Large, WavLM-Large, and Wav2Vec2-XLSR-300m. These models consist of a CNN encoder followed by 24 Transformer layers, producing a 1024-dimensional output representation, with a total of 316 million parameters. For back-end classification, we compare our ResNet-based model against three architectures: AASIST [14], Conformer [3], and MHFA [4].

4. Results and Analysis

Our experimental framework systematically benchmarks the combination between various Self-Supervised Learning (SSL) front-ends and back-end classifiers. We begin by evaluating four distinct back-ends paired with an XLSR feature extractor (Sec. 4.1), followed by qualitative results of the learned embedding space via visualization (Sec. 4.2). Subsequently, we as-



(a) *Training Case 1.* (b) *Training Case 2.* (c) *Training Case 3.*

Figure 2: *t-SNE visualization of learned embeddings of XLSR-ResNet model in three training cases.*

Table 4: *Model performance (% EER) of the XLSR-ResNet model in training case 1 and case 2, evaluated on ASVspoof 5 eval break down: codecs (C01-11). The third column ("-") represents performance on non-codec audio, while "pooled" indicates the average performance across all categories.*

	pooled	-	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11
Training Case 1	4.72	0.95	2.70	3.22	3.15	11.87	1.03	1.33	14.52	5.24	3.26	4.30	1.27
Training Case 2	12.34	0.53	8.58	13.68	6.95	27.11	3.33	2.52	32.32	10.24	9.77	11.66	1.07

sess the compatibility of alternative SSL architectures with our ResNet back-end (Sec. 4.3). The study concludes with an investigation into the model's cross-linguistic transferability and generalization (Sec. 4.4).

4.1. Comparison of Back-end Classifiers

We benchmark four back-end architectures (i.e. AASIST, Conformer, MHFA, and our proposed ResNet) across three distinct training cases (see Sec. 3.1). As summarized in Table 2, the ResNet back-end gives the best average performance, surpassing the competitive MHFA baseline with relative EER reductions of 10.4%, 4.1%, and 9.2% across Cases 1, 2, and 3, respectively.

While multi-corpus training generally scales model efficacy, the results reveal a nuanced sensitivity to data composition. On the Wild dataset, ResNet's EER decreases by 48% (Case 2) and 69% (Case 3) relative to Case 1. Similarly, the inclusion of the MLAAD dataset (Case 2) precipitously reduces EER from 4.38% to 0.17% for the Fake-or-Real evaluation set. Conversely, a notable performance regression occurs on the ASVspoof 5 evaluation set following the introduction of external training data, where the EER increases from 4.73% (Case 1) to 12.37% (Case 2). To investigate this anomaly, we decouple the results by audio type (Table 4). The degradation is isolated to codec-processed audio (compressed/transmitted), whereas performance on non-codec audio actually improves (EER declines from 0.95% to 0.53%). This suggests a significant domain mismatch or dataset bias between ASVspoof 5 and MLAAD with respect to channel effects.

4.2. Visualisation

To further diagnose the bias issue in the previous section, we visualize the ResNet embeddings using t-SNE (Fig. 2). In Case 1 (trained solely on ASVspoof 5), spoofed evaluation samples (green) exhibit a fragmented, non-compact distribution, with several samples encroaching on the bonafide region (red). In Case 2, while the addition of MLAAD data improves the alignment of MLAAD spoof samples with the ASVspoof 5 spoof cluster, it simultaneously pulls ASVspoof 5 bonafide embeddings toward the spoof region. This shift introduces representation confusion, explaining the performance drop in Case 2. Furthermore, the distinct clustering by dataset source, rather than just class label, confirms the presence of dataset bias.

4.3. Comparison of SSL-based Feature Extractors

We evaluate four SSL front-ends integrated with our ResNet back-end (Table 3). XLSR demonstrates superior efficacy, achieving a benchmark EER of 2.02% on the Wild dataset. This robustness is attributed to its extensive pre-training on 436,000 hours of diverse, multilingual audio, which facilitates a highly generalized representation of global speech variability. Although Wav2vec 2.0 and HuBERT exhibit higher average EERs compared to WavLM and XLSR, they remain remarkably competitive on the ASVspoof 5 evaluation set. We hypothesize that this contextualized success stems from a pre-training bias: both models were trained on the Libri-Light corpus, which was also utilized to develop several ASVspoof 5 attack models.

Table 5: *Performance on non-target languages (HABLA: Spanish dataset; CFAD: Chinese dataset) in training case 1.*

Method	HABLA		CFAD	
	EER	minDCF	EER	minDCF
XLSR + AASIST	18.43	0.33	30.16	0.66
XLSR + Conformer	13.58	0.28	23.15	0.55
XLSR + MHFA	19.14	0.34	22.14	0.48
XLSR + ResNet	17.90	0.32	27.23	0.56

Table 6: *Performance on other languages in training case 2: ASVspoof 5 training and MLAAD-v3 dataset (MLAAD-v3 includes spoofed Spanish audio but excludes Chinese)*

Method	HABLA		CFAD	
	EER	minDCF	EER	minDCF
XLSR + AASIST	6.33	0.16	26.88	0.64
XLSR + Conformer	4.39	0.11	27.11	0.49
XLSR + MHFA	5.53	0.13	23.23	0.47
XLSR + ResNet	2.98	0.07	23.54	0.46

4.4. Cross-linguistic Generalization

We evaluate the cross-linguistic robustness of our models using the Spanish (HABLA) and Chinese (CFAD) datasets (Tables 5 and 6). In Case 1, under monolingual supervision on English data (ASVspoof 5), models exhibit a substantial performance deficit on non-English evaluation sets. The Conformer architecture yields the most competitive baseline, with EERs of 13.58% on HABLA and 23.15% on CFAD. This underscores that models optimized exclusively for English feature spaces fail to encapsulate the phonetic and prosodic signatures of spoofing attacks in diverse linguistic contexts, manifesting a significant linguistic domain gap. The introduction of the MLAAD dataset (Case 2), which comprises 23 languages including Spanish but excluding Chinese, yielded divergent results. On the Spanish HABLA set, EERs decline precipitously. Despite the sparsity of Spanish spoofed audio in MLAAD (approximately 8 hours), our ResNet achieves a 2.98% EER. This demonstrates that even minimal exposure to target-language variance facilitates effective cross-linguistic alignment. Conversely, performance on the Chinese CFAD set remains invariant. While our results highlight the necessity of including target-language data during training, further experiments across a broader range of languages are required to generalize this hypothesis.

5. Conclusions

In this study, we conduct a comprehensive evaluation of spoofing detection models, focusing on the integration of SSL-based feature extractors and back-end classifiers across various training and evaluation scenarios. Our findings show the ResNet classifier to be the most effective. It achieves consistently the lowest average EER. Our analysis of SSL-based feature extractors reveals substantial performance variations, with XLSR standing out as the most effective due to extensive pre-training using diverse datasets. This underscores the critical role of large-scale, diverse pre-training in enhancing feature representations for the spoofing detection task. Additionally, our multi-corpus training experiments expose significant dataset biases, particularly for the ASVspoof 5 evaluation set, which highlights the need for robust bias mitigation strategies to improve generalization. Furthermore, our cross-linguistic evaluation for Spanish (HABLA dataset) and Chinese (CFAD dataset) languages demonstrates that the incorporation of even a modest amount of target-language data (approximately 8 hours) yields marked improvements in detection performance. This finding emphasizes the importance of language-specific adaptation, particularly for underrepresented languages, to spoofing detection robustness in multilingual contexts.

6. References

[1] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *ASVspoof*, 2024.

[2] J. Jung, H. Heo, H. Tak, H. Shim, J. S. Chung, B. Lee, H. Yu, and N. Evans, "AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP*, 2022.

[3] E. Rosello, A. G. Alanís, A. M. Gomez, and A. M. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Interspeech*, 2023.

[4] T. Stourbe, V. Miara, T. Lepage, and R. Dehak, "Exploring WavLM back-ends for speech spoofing and deepfake detection," in *ASVspoof 2024*, 2024.

[5] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised xls-r and sls classifier." Association for Computing Machinery, 2024.

[6] Y. Xiao and R. K. Das, "Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, 2025.

[7] 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 *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[8] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11*. IEEE, 2021, pp. 6369–6373.

[9] A.-T. Dao, M. Rouvier, and D. Matrouf, "ASVspoof 5 challenge: advanced resnet architectures for robust voice spoofing detection," in *ASVspoof*, 2024.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] W.-N. Hsu, B. Bolte, Y. Y. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," in *CVPR*, 2021.

[12] S. Chen, C. Wu, Z. Yang, T. Yao, Y. Zhang, S. Liu, J. Li, M. Zhou, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[13] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022.

[14] T. Hemlata, T. Massimiliano, W. Xin, J. Jee-weon, Y. Junichi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Odyssey*, 2022.

[15] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2023.

[16] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

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

[18] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[19] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021*, 2021.

[20] R. Reimao and V. 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*, C. Burileanu and H. Teodorescu, Eds., 2019.

[21] P. A. T. Flórez, R. Manrique, and B. P. Nunes, "HABLA: A dataset of latin american spanish accents for voice anti-spoofing," in *Interspeech*. ISCA, 2023, pp. 1963–1967.

[22] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A chinese dataset for fake audio detection," *Speech Commun.*, vol. 164, p. 103122, 2024.

[23] J. M. Martín-Doñas, A. Álvarez, E. Rosello, A. M. Gomez, and A. M. Peinado, "Exploring Self-supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection," in *Interspeech*, 2024.

[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *ArXiv*, 2015.

[25] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.