# The Multilingual Scaling Plateau in Speech Foundation Models for Anti-Spoofing

Anonymous ACL submission

## Abstract

Speech foundation models (SFMs) have become the dominant backbone for modern speech deepfake detection systems. However, despite rapid scaling in multilingual coverage and model size, it remains unclear how multilingual pretraining affects robustness and out-of-domain (OOD) generalization for spoofing detection. In this work, we present a large-scale systematic study of 20 SFMs spanning English-only, moderately multilingual, and massively multilingual pretraining regimes across model scales. We evaluate all models across 22 evaluation conditions covering in-domain benchmarks, vocoder artifacts, diffusion and flow-matching synthesis, in-the-wild recordings, and multilingual spoofing datasets. Our results reveal three key findings. First, multilingual pretraining SFMs consistently improves robustness over English-only pretraining, even on predominantly English evaluation sets. Second, we observe diminishing returns from extreme multilingual scaling: medium-scale multilingual pretrained SFMs consistently provides the best trade-off between in-domain performance and OOD generalization. Third, model scaling provides complementary gains but does not override the importance of pretraining data quality. Overall, `XLS-R-128` achieves the most consistent performance across evaluation conditions, establishing a strong robustness-performance trade-off for deepfake detection.

## 1 Introduction

Recent advances in neural speech synthesis have enabled synthetic speech systems capable of generating highly natural and speaker-consistent audio (Tan et al., 2021). Modern text-to-speech (TTS) and voice conversion (VC) systems based on diffusion models, flow matching, and large-scale neural vocoders can now produce speech that is often difficult to distinguish from genuine recordings (O'Shaughnessy, 2026). While these advances benefit accessibility, dubbing, and conversational AI, they also introduce serious risks including identity fraud, impersonation (Folorunsho and Boamah, 2025), and attacks against automatic speaker verification (ASV) systems (Todisco et al., 2019; Yamagishi et al., 2021; Wang et al., 2024).

To address these threats, the anti-spoofing community has increasingly adopted self-supervised speech foundation models (SFMs) as universal acoustic front-ends for deepfake detection. Models such as `Wav2Vec 2.0` (Baevski et al., 2020), `WavLM` (Chen et al., 2022a), `XLS-R` (Babu et al., 2022) provide rich speech representations learned from large-scale unlabeled speech corpora. Existing works (Tran et al., 2024, 2025a,b; Yang et al., 2025a,b; Tran et al., 2025c, 2026; Dat et al., 2026; Truong et al., 2026; Yang et al., 2026; Zhao et al., 2026; Xuan et al., 2026; El Kheir et al., 2026; Buera et al., 2026; Ali et al., 2026; Dowerah et al., 2026) have demonstrated that such representations significantly outperform traditional handcrafted features (Lei and Lopez, 2009; Todisco et al., 2017).

Despite these advances, the role of multilingual SFMs when fine-tuned for deepfake detection remains poorly understood. A prevailing assumption in speech representation learning is that larger and more multilingual pretraining corpora consistently improve downstream robustness (Pu et al., 2021). While this trend has largely held by fine-tuning for automatic speech recognition (ASR) (Pratap et al., 2020a, 2024; Omnilingual et al., 2025), it is unclear whether the same scaling behavior can be applied for anti-spoofing, where the objective is not semantic understanding but the detection of subtle synthesis artifacts.

In this work, we conduct a systematic empirical analysis of monolingual and multilingual scaling in SFMs when fine-tuned for anti-spoofing. We benchmark 20 representative models spanning English-only, moderately multilingual, and massively multilingual pretraining regimes under a unified experimental protocol. Our evaluation cov-

ers 22 in-domain and OOD in-the-wild conditions including diffusion-based synthesis, vocoder artifacts, and multilingual spoofing datasets. Our study leads to several important observations. Multilingual pretraining consistently improves robustness compared to English-only pretraining, even on English benchmarks. However, we do not observe consistent gains from extreme multilingual scaling beyond medium-scale multilingual pretraining. Instead, SFMs pretrained on 128 languages consistently achieve the strongest robustness-performance trade-off across evaluations. Our contributions are as follows:

1. A comparison of 20 SFMs ranging from monolingual to 1,600+ language pretraining when fine-tuned for speech deepfake detection.

2. Evidence for a linguistic diversity plateau: performance improves consistently from English-only to 128 pretraining languages but plateaus or degrades beyond this threshold.

3. We evaluate across multiple spoofing datasets, including in-domain and OOD.

## 2 Related works

### 2.1 Speech deepfake synthesis

Modern TTS systems generate speech through a multi-stage pipeline: linguistic representations are mapped to acoustic features such as mel-spectrograms, which are then synthesized into waveforms by neural vocoders (O'Shaughnessy, 2026). VC systems preserve linguistic content while transforming speaker identity using reference speech (Tan et al., 2021). With large-scale training data and powerful architectures, both TTS and VC systems now produce highly natural speech that is perceptually indistinguishable from genuine recordings, enabling increasingly convincing deepfakes. Notably, diffusion-based and flow-matching-based TTS methods (Eskimez et al., 2024; Ju et al., 2024; Chen et al., 2025) represent the latest frontier, generating speech with high fidelity and posing new challenges to existing detectors.

### 2.2 Anti-spoofing countermeasures

Early anti-spoofing systems relied on handcrafted acoustic descriptors such as MFCC, LFCC, and CQCC features (Lei and Lopez, 2009; Todisco et al., 2017). Recent work has increasingly adopted self-supervised SFMs as universal front-end encoders. Models such as `Wav2Vec 2.0` (Baevski et al., 2020), `WavLM` (Chen et al., 2022a), `XLS-R` (Babu et al., 2022) have demonstrated strong performance across speech processing tasks including speaker verification, and speech recognition (wen Yang et al., 2021). The anti-spoofing community has converged on SFM-based feature extractors as the dominant paradigm (Ali et al., 2026; Dowerah et al., 2026). Classifier architectures have diversified from LSTMs (Guan et al., 2025) and graph-based models (Tak et al., 2022b; Yang et al., 2025b,a) to transformer-based designs (Rosello et al., 2023; Truong et al., 2024; Li et al., 2024; Hao et al., 2025; Dat and Dat, 2025; Kim et al., 2025; Phuong et al., 2025; Tran et al., 2025a; Truong et al., 2025b,a). To reduce the computational cost of full self-attention, state-space models (Tran et al., 2025b; Xiao and Das, 2025; Xuan et al., 2025) and gated multilayer perceptrons (Tran et al., 2025c) have been proposed as efficient alternatives. Despite these advances, generalization to unseen attacks and out-of-domain data remains an open problem (Müller et al., 2022).

### 2.3 Multilingual speech foundation models

Large-scale multilingual pretraining has emerged as a dominant paradigm in speech representation learning (Shi et al., 2023). Models such as `XLS-R` (Conneau et al., 2021; Babu et al., 2022), `MMS` (Pratap et al., 2024), and `Omnilingual` (Omnilingual et al., 2025) scale multilingual speech learning to hundreds or thousands of languages. These models are generally motivated by the hypothesis that linguistic diversity improves representation robustness and cross-lingual transferability (Pu et al., 2021; Pratap et al., 2020a, 2024; Omnilingual et al., 2025). However, whether extreme multilingual scaling benefits deepfake detection remains unclear. Unlike ASR, spoofing detection primarily depends on identifying synthesis artifacts rather than semantic or phonetic content. This motivates a systematic investigation of multilingual scaling behavior for speech deepfake detection.

### 2.4 SFM-based deepfake detection

Kulkarni et al. (2024) utilized English-only SFMs to evaluate their performance in spoofing detection. Their findings indicate that while fusion approaches yield superior results, they necessitate training multiple models independently prior to integration. Most contemporary systems employ `XLS-R-128` (§ 2.2) as a feature extractor, demonstrating the advantages of multilingual pretraining over English-

only alternatives. Furthermore, Tran et al. (2025b) utilized MMS, while Phuong et al. (2025) investigated the BASE-sized mHuBERT-147 (Zanon Boito et al., 2024). Although Ali et al. (2026) recently designed an SFM-based spoofing benchmark, the majority of existing SFMs remain English-only and are typically evaluated only on a limited set of OOD datasets. Specifically, El Kheir et al. (2025) performed a layer-wise investigation into the generalizability of English-only SFMs. To bridge this gap, we provide a comprehensive benchmark of SFMs, spanning from monolingual English to multilingual models, with scales ranging from BASE (0.1B) to LARGE (0.3B) and XL (1B).

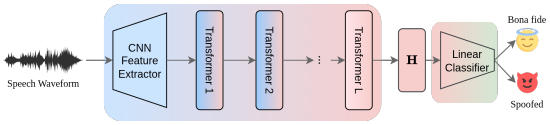


Figure 1: Overview of the framework with speech foundation model backbone (CNNs and $L$ Transformer layers) followed by a linear classifier for anti-spoofing.

## 3 Method

### 3.1 Architecture

As shown in Figure 1, a waveform $\mathbf{x}$ is mapped to latents $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ via a CNN feature extractor, then processed by $L$ Transformer layers. Let $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\} \in \mathbb{R}^{T \times D}$ denote the hidden states from the final layer, where $T$ is the number of frames and $D$ is the hidden dimension. We apply temporal average pooling to get the utterance representation $\bar{\mathbf{H}}$. A linear head computes the binary logits $\hat{\mathbf{y}} = c(\bar{\mathbf{H}})$ (bona fide vs. spoof).

### 3.2 Speech foundation models

To comprehensively investigate the impact of pretraining paradigms on speech deepfake detection, we evaluate 20 representative SFMs. These models span diverse self-supervised learning (SSL) objectives, training corpora, architectural scales, and linguistic coverages. We categorize the selected SFMs along two primary design axes: linguistic scope (English-only vs. multilingual) and parameter scale ranging from BASE (0.1B) to LARGE (0.3B) and XL (1B). A detailed description of these models is summarized in Table 1 and Appendix A.

#### 3.2.1 English-only models

We consider four prominent English pretrained SFMs at the BASE and LARGE scales:

**Wav2Vec 2.0.** A contrastive learning framework that minimizes a contrastive loss over masked latent acoustic representations (Baevski et al., 2020).

**HuBERT.** An offline clustering and masked language modeling approach that forces the network to predict discrete hidden units (Hsu et al., 2021).

**UniSpeech-SAT.** A SFM incorporating a speaker-aware pretraining strategy via utterance-level contrastive loss to retain speaker identity (Chen et al., 2022b).

**WavLM.** An extension of the masked speech denoising and generation framework, which utilizes gated relative position embeddings and an overlapping speaker mixing strategy to improve robustness for complex auditory scenes (Chen et al., 2022a).

For both UniSpeech-SAT and WavLM, we evaluate variants pretrained on standard LibriSpeech (Panayotov et al., 2015) (BASE) as well as expanded setups utilizing larger English corpora including VoxPopuli (Wang et al., 2021) and GigaSpeech (Chen et al., 2021) (BASE+).

#### 3.2.2 Multilingual models

To evaluate cross-lingual generalization and the effect of massive multilingual pretraining, we integrate several cross-lingual SFMs across all scales:

**Multilingual variants of Wav2Vec2/HuBERT.** This includes **Wav2Vec2-Vox**, pretrained on the multilingual VoxPopuli corpus (Wang et al., 2021), another variant of 56 thousand hours of speech covering 53 (**XLSR-53**) languages (Conneau et al., 2021), and **mHuBERT-147** (Zanon Boito et al., 2024), a compact architecture capturing 147 languages via hidden-unit prediction.

**XLS-R.** A large-scale multilingual extension of Wav2Vec 2.0 trained on up to 436,000 hours of speech 128 (**XLS-R-128**) languages, evaluated at both the LARGE and XL (1B parameters) scales.

**MMS.** (Pratap et al., 2024): The massively multilingual SFM, expanding coverage to over 1,000, scaled up to XL.

**Omnilingual.** (Omnilingual et al., 2025): A state-of-the-art SFM optimized for ultra-diverse linguistic representations encompassing over 1,600 languages, evaluated at the LARGE and XL capacities.

| Model | Variant | Pretraining corpus (Hours) | Languages | Parameters (B) | Network architecture |
|---|---|---|---|---|---|
| Wav2Vec 2.0 | BASE / LARGE | LibriSpeech (960h) | English | 0.1B / 0.3B | 7 Conv, 12 / 24 Trans |
| | VoxPopuli BASE / LARGE | VoxPopuli (100k h) | 23 languages | 0.3B | 7 Conv, 24 Trans |
| | XLSR-53 LARGE | Multilingual LibriSpeech, CommonVoice, Babel (56k h) | 53 languages | 0.3B | 7 Conv, 24 Trans |
| HuBERT | BASE / LARGE | LibriSpeech (960h) / Libri-Light (60k h) | English | 0.1B / 0.3B | 7 Conv, 12 / 24 Trans |
| | mHuBERT-147 BASE | Open-license data (90k h) | 147 languages | 0.1B | 7 Conv, 12 Trans |
| UniSpeech-SAT | BASE (+) / LARGE | LibriSpeech (960h) / Librivox, VoxPopuli, GigaSpeech (94k h) | English | 0.1B / 0.3B | 7 Conv, 12 / 24 Trans |
| WavLM | BASE (+) / LARGE | LibriSpeech (960h) / Librivox, VoxPopuli, GigaSpeech (94k h) | English | 0.1B / 0.3B | 7 Conv, 12 / 24 Trans |
| XLS-R | LARGE / XL | Multilingual LibriSpeech, CommonVoice, Babel, VoxPopuli, VoxLingua107 (436k h) | 128 languages | 0.3B / 1B | 7 Conv, 24 / 48 Trans |
| MMS | LARGE / XL | Bible.is, VoxPopuli, etc. (491k h) | 1,406 languages | 0.3B / 1B | 7 Conv, 24 / 48 Trans |
| Omnilingual | LARGE / XL | Public + Internal speech pool (4.3M h) | 1,600+ languages | 0.3B / 1B | 7 Conv, 24 / 48 Trans |

Table 1: Comparison of representative speech foundation models used in this study in terms of pretraining corpus, multilingual coverage, scale, and network architecture.

# 4 Experiments

**Training and development.** We train all models using the ASVspoof 2019 logical access (19LA) training partition and select the best configurations based on the 19LA development set, which features a disjoint set of speakers. The synthetic speech in 19LA is generated via TTS and VC techniques.

**In-domain evaluation.** To assess in-domain performance, we test the models on the 19LA evaluation set, alongside the ASVspoof 2021 logical access (21LA) and deepfake (21DF) subsets. While 21LA introduces codec and transmission distortions to the 19LA setup, 21DF incorporates diverse lossy compression artifacts, offering a more realistic evaluation within the standard ASVspoof framework.

**Out-of-domain evaluation.** We evaluate model generalization and robustness against distribution shifts using a diverse suite of OOD datasets. These include in-the-wild recordings, diffusion- and vocoder-based synthesis, cross-lingual samples, and large-scale multilingual benchmarks. Comprehensive descriptions and statistics for these datasets are available in Appendix B.

**Performance metric.** In line with prior works, we use the equal error rate (EER) as our primary metric where a lower EER signifies a better system. Detailed definitions are provided in Appendix C.

**Implementation details.** We fine-tune pretrained SFMs from HuggingFace. Audios in each batch are padded to match the longest duration. To address the class imbalance in 19LA, we use a weighted cross-entropy loss that gives more weight to the bona fide class. We also apply RawBoost data augmentation (Tak et al., 2022a) to add various types of noise. More information can be found in Appendix D.

| Size | Model | Lang. | Avg. |
|---|---|---|---|
| BASE | Wav2Vec2 | EN | 15.40 |
| | HuBERT | EN | 15.63 |
| | UniSpeechSAT | EN | 13.95 |
| | UniSpeechSAT+ | EN | 11.49 |
| | WavLM | EN | 15.50 |
| | WavLM+ | EN | 13.11 |
| | Wav2Vec2 Vox. | Multi | 14.55 |
| | mHuBERT-147 | Multi | 9.46 |
| LARGE | Wav2Vec2 | EN | 15.82 |
| | HuBERT | EN | 11.79 |
| | UniSpeechSAT | EN | 22.06 |
| | WavLM | EN | 23.19 |
| | Wav2Vec2 Vox. | Multi | 19.25 |
| | XLS-R-53 | Multi | 11.59 |
| | XLS-R-128 | Multi | 7.34 |
| | MMS | Multi | 7.20 |
| | Omnilingual | Multi | 8.42 |
| XL | XLS-R-128 | Multi | 5.36 |
| | MMS | Multi | 6.14 |
| | Omnilingual | Multi | 6.24 |

Table 2: Average EER (%, ↓) across all evaluated SFMs. Models are grouped by scale and pretraining language scope (EN = English-only, Multi = multilingual). The best model per language scope is highlighted in green.

# 5 Results and analysis

In this section, we systematically analyze the downstream performance of SFMs. The detailed results are provided in the Appendix, with full breakdowns across English evaluation domains in Table 5 and multilingual benchmarks in Table 6.

## 5.1 Monolingual scaling collapse vs. multilingual regularization

Our finding indicates that multilingual pretraining acts as a robust regularizer for speech deepfake detection (Tab. 2). Notably, it consistently outperforms monolingual SFMs even when evaluated exclusively on English target domains (Fig. 2, 3). At the BASE scale, the top-performing English-only SFM (UniSpeechSAT+, achieving an 11.49% average EER) is significantly outperformed by mHuBERT-147 (9.46% average EER). As illustrated in Figure 2 (b), the best multilingual models consistently achieves lower EER than their monolingual SFMs. This confirms that training on diverse, cross-lingual data helps the model generalize better.

This performance divergence expands dramatically at the LARGE scale. The best-performing

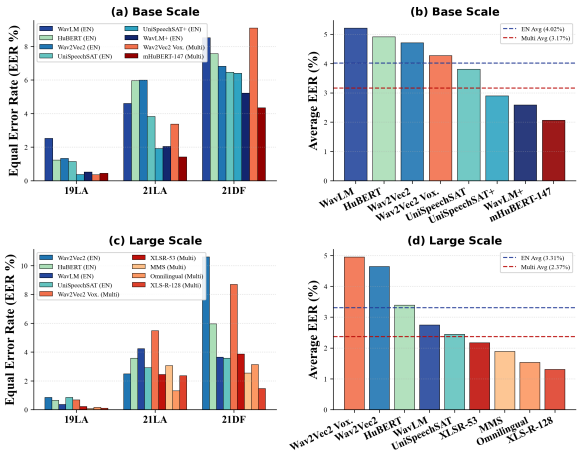


Figure 2: In-domain performance (EER %↓) (19LA, 21LA, 21DF) of speech foundation models (SFMs) scaled across BASE and LARGE configurations. Left panels (a, c) report EER for individual dataset, categorized by English-only (cool colors) and multilingual (warm colors) SFMs. Right panels (b, d) show the average EER.

English-only baseline (HuBERT: 11.79% average EER) is heavily outperformed by XLS-R-128 (7.34% average EER). As shown in Figure 3, the cross-dataset performance clearly demonstrates the generalization benefits of language diversification.

However, we observe a severe performance degradation when scaling model parameters within a single language domain. LARGE English-only models structurally underperform their corresponding BASE architectures; specifically, WavLM LARGE (23.19% average EER) and UniSpeechSAT LARGE (22.06% average EER) degrade significantly relative to WavLM BASE (15.50% EER) and UniSpeechSAT BASE (13.95% EER).

As evidenced by the distinct error spikes across individual datasets in Figure 3, parameter scaling in a monolingual pretraining regime causes the encoder to overfit to language-specific features. This introduces severe brittleness under OOD distribution shifts. Conversely, multilingual pretraining acts as a regularizer, forcing the model to learn robust, and domain-invariant features.

### 5.2 The linguistic diversity plateau and capacity scaling

Another finding of our analysis is that downstream speech deepfake detection performance does not monotonically improve with increasing multilingual coverage during self-supervised pretraining. Instead, we observe a non-linear relationship in which expanding language diversity beyond a certain point yields diminishing returns and, in some cases, degraded performance. Within the LARGE multilingual model family (Table 2), EERs decrease substantially when scaling from 53 to 128 languages, but plateau or worsen when moving to models pretrained on more than 1,000 languages.

This trend is consistently reflected across multiple OOD evaluation sets in Figure 3. On the real-world celebrity SC, XLS-R-128 achieves a 19.57% EER, outperforming both MMS (20.50%) and Omnilingual (20.22%). Similarly, on ITW, XLS-R-128 obtains a 4.69% EER, compared to 7.08% for MMS and 5.96% for Omnilingual.

### 5.3 Optimal multilingual scale

Our results suggest that multilingual diversity benefits speech deepfake detection only up to a certain regime. Scaling from 53 languages (XLSR-53) to 128 languages (XLS-R-128) substantially improves robustness, likely because the model gains broader phonetic coverage while still maintaining sufficiently large per-language speech representations. Importantly, this transition is accompanied not only by increased language diversity, but also by a nearly tenfold increase in total pretraining data volume, from 56k to nearly 500k hours of speech.

However, further scaling toward hyper-multilingual settings (MMS and Omnilingual) does not continue this trend. Although these models cover more than one thousand languages, the underlying data distributions become substantially more long-tailed, with many languages represented by comparatively limited quantities of speech. Consequently, the SSL must learn increasingly language-invariant representations across highly heterogeneous acoustic and linguistic conditions.

For semantic speech tasks such as ASR, such invariance is highly beneficial. In contrast, speech deepfake detection relies heavily on preserving subtle low-level acoustic artifacts, including synthesis imperfections, vocoder traces, and recording-channel signatures. Excessive pressure toward broad multilingual invariance may therefore reduce sensitivity to these fine-grained cues, leading to the observed multilingual scaling plateau.

### 5.4 Model capacity as an independent scaling factor

Model scaling remains beneficial even when language diversity is held constant. Expanding models from LARGE to XL configurations consistently improves performance across all multilingual SFMs.

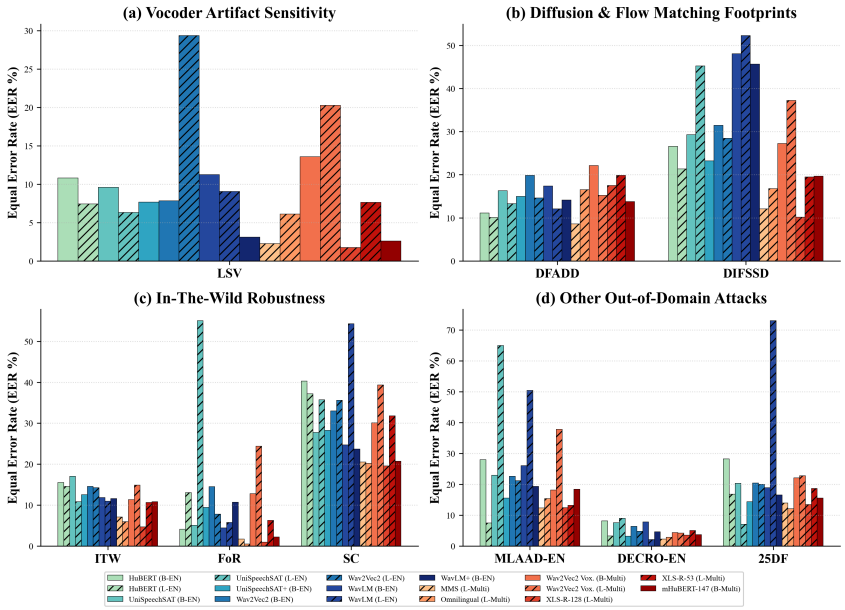


Figure 3: Out-of-domain (OOD) performance (EER %↓). Panels correspond to: (a) Vocoder artifacts (LSV); (b) Diffusion and flow matching footprints (DFADD, DIFSSD); (c) In-the-wild conditions (ITW, FoR, SC); and (d) Other OOD attacks (MLAAD-EN, DECRO-EN, 25DF). SFMs are separated by scope (cool / warm) and scale (hatched / solid).

For example, XLS-R-128 improves from 7.34% to 5.36% average EER, corresponding to a 27.0% relative error reduction. Likewise, MMS improves from 7.20% to 6.14%, while Omnilingual decreases from 8.42% to 6.24%. However, increasing model capacity alone does not eliminate the multilingual scaling plateau. Even at the XL scale, XLS-R-128 XL achieves the strongest overall performance (5.36% average EER), outperforming both MMS XL (6.14%) and Omnilingual XL (6.24%) despite being pretrained on substantially fewer languages. These results suggest that pretraining data composition plays a critical role in downstream speech deepfake detection, and that increasing parameter count cannot fully compensate for suboptimal pretraining distributions.

This parameter capacity limitation is further illustrated by recent explorations into extreme scale SFMs for anti-spoofing. When moving beyond medium-scale models to massive scaled configurations–such as evaluating the performance leap from the 1-billion parameter XLS-R-128 (XL) to the 2-billion parameter XLS-R-128 (XXL)–empirical evidence demonstrates clear diminishing returns (Ge et al., 2025). While the XXL size provides small improvements in OOD robustness, the shrinking gap between models proves that making an architecture larger eventually hits a limit.

### 5.5 Fine-grained dataset analysis

**Clean vs. compression shifts.** While clean, studio-quality recordings (19LA) result in near-zero EER for almost all models, adding compressed audio or channel changes (21LA, 21DF) severely hurts monolingual systems (Figure 2. For example, Wav2Vec 2.0 BASE increases from 1.33% to 6.81% EER on 21DF. This shows that monolingual models often mistake compression shifts.

**Vocoder artifacts.** XLS-R-128 LARGE achieves a remarkable 1.75% EER, far better than monolingual models (WavLM LARGE: 9.06%; HuBERT LARGE: 7.42%). Multilingual representations appear to better capture vocoder artifacts.

**Diffusion/flow-matching footprints.** These datasets represent the current frontier for anti-spoofing. Monolingual BASE models collapse on DIFFSSD (WavLM: 48.05%). XLS-R-128 LARGE achieves 10.16% EER, confirming that LARGE SFMs best handle next-generation synthesis.

**In-the-wild conditions.** These attacks are universally challenging, even the best model (Omnilingual XL: 15.09%) yields high EER on SC, highlighting the need for in-the-wild or domain-adapted training data. For ITW, Omnilingual XL still achieves the best EER of 4.22%, followed closely by MMS XL (4.92%) and XLS-R-128 XL (5.73%), suggesting that broader language cover-

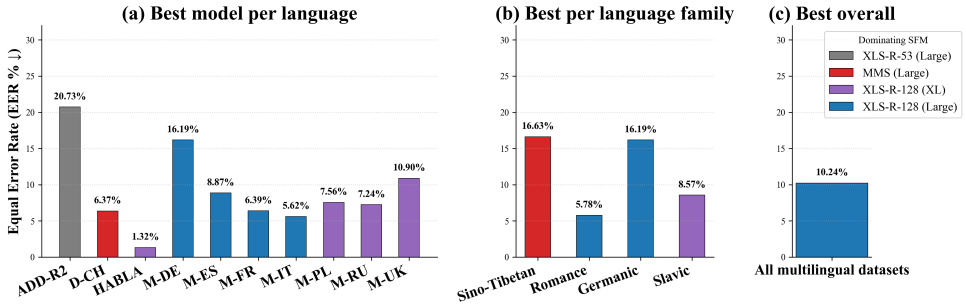


Figure 4: Analysis of multilingual SFMs across cross-lingual anti-spoofing benchmarks. (a) Best-performing model for each language subset. (b) Best-performing model aggregated by language family. (c) Overall strongest multilingual SFM averaged across all multilingual datasets.

age may marginally benefit detection of highly diverse in-the-wild synthesis systems.

**Other out-of-domain attacks.** While DECRO-EN is highly tractable for most multilingual systems, MLAAD-EN and 25DF reveal a degraded performance among large monolingual SFMs. Specifically, WavLM LARGE degrades heavily to 50.41% EER on MLAAD-EN and 73.01% EER on 25DF, highlighting their severe brittleness under distribution shifts. In contrast, multilingual architectures remain robustly controlled across both tasks. XLS-R-128 XL achieves the top performance on MLAAD-EN (6.72% EER), while MMS XL leads on the adversarial conditions of 25DF (10.24% EER), likely benefiting from its superior speaker and channel diversity.

### 5.6 Cross-lingual and cross-accent generalization

Figure 4 summarizes the robustness of SFMs under cross-lingual and cross-accent distribution shifts across multilingual benchmarks, including MLAAD, HABLA, D-CH, and ADD-R2.

**Monolingual pretraining limits cross-lingual transfer.** English-only SFMs consistently degrade when evaluated on unseen languages. For example, WavLM BASE reaches near-random performance on several MLAAD subsets, including German (M-DE: 48.21% EER), Spanish (M-ES: 47.36%), and French (M-FR: 48.09%). Scaling does not resolve this issue: WavLM LARGE still performs poorly on Russian (M-RU: 68.62%), while UniSpeechSAT LARGE degrades substantially on Italian (M-IT: 66.30%) and Polish (M-PL: 67.43%). These results indicate that increasing model capacity without sufficient multilingual diversity does not produce robust language-invariant representations.

**Large multilingual SFMs provide substantially stronger transfer.** In contrast, multilingual pretraining consistently improves cross-lingual robustness. XLS-R-128 LARGE achieves 16.19% EER on German, 8.87% on Spanish, 6.39% on French, and 5.62% on Italian, showing improvements over monolingual baselines. Figure 4 (a) further shows that multilingual XLS-R variants dominate most European language subsets, particularly within Romance and Germanic families.

**Cross-accent robustness emerges at larger multilingual scales.** The strongest result is observed on HABLA, where XLS-R-128 XL achieves only 1.32% EER on Latin American Spanish accents. This suggests that large-scale multilingual pretraining not only improves language transfer, but also enhances robustness to accentual and phonetic variability within the same language family.

**Multilingual scaling is beneficial but not universally dominant.** Despite the overall superiority of highly multilingual models, Figure 4 (b) reveals important language-family-specific exceptions. On the Mandarin benchmark ADD-R2, XLS-R-53 LARGE (20.73% EER) outperforms both XLS-R-128 LARGE (22.78%) and MMS LARGE (26.90%). Similarly, on the Chinese D-CH benchmark, MMS LARGE achieves the best result (6.37% EER), outperforming XLS-R-128 LARGE (13.08%) and Omnilingual LARGE (11.76%). These findings suggest that robustness is not determined solely by the number of pretraining languages, but also by the distributional characteristics and linguistic composition of the pretraining corpus.

**Hierarchical robustness trends across language families.** The analysis in Figure 4 highlights a clear hierarchical pattern. Romance languages exhibit the lowest best-case EERs (5.78%), fol-

lowed by Slavic languages (8.57%), while Germanic languages remain substantially more challenging (16.19%). Sino-Tibetan benchmarks remain the most difficult overall, with best-case performance still above 16% EER. Furthermore, Figure 4 (c) shows that XLS-R-128 LARGE achieves the strongest average robustness across all multilingual datasets (10.24% EER), indicating that large-scale multilingual pretraining provides the most consistent cross-lingual generalization overall.

Overall, these results demonstrate that multilingual diversity and model scaling interact synergistically, but their effectiveness depends strongly on language-family coverage and pretraining distributional alignment. Large multilingual SFMs substantially improve robustness to cross-lingual and cross-accent anti-spoofing, although important language-specific failure modes still remain.

### 5.7 Upstream SSL pretraining objectives and architectural effects

To isolate the impact of upstream SSL objectives, we analyze the English-only BASE-scale models, where architectural capacity remains relatively controlled and performance differences primarily reflect pretraining strategy. Models incorporating speaker-aware or acoustically enriched objectives consistently outperform their vanilla counterparts. UniSpeechSAT+ achieves an average EER of 11.49%, substantially improving over the baseline UniSpeechSAT (13.95%), while WavLM+ reaches 13.11% compared with 15.50% for standard WavLM. These results suggest that speaker-aware contrastive learning and multi-speaker utterance mixing encourage the encoder to preserve fine-grained speaker and acoustic characteristics that remain informative for detecting artifacts.

Clustering-based hidden-unit prediction also proves highly effective. Despite being trained on fewer languages than massively multilingual models, mHuBERT-147 achieves a strong macro-average EER of 9.46%, outperforming several larger multilingual alternatives at the same scale. This indicates that HuBERT-style discrete unit prediction captures phonetic and acoustic boundary information particularly useful for spoofing detection. Consequently, the results demonstrate that upstream SSL objectives strongly influence downstream deepfake detection robustness. Architectures that explicitly model speaker variability or discrete acoustic structure produce substantially more transferable representations than only masked prediction objectives.

## 6 Discussion

Our large-scale evaluation demonstrates that the scaling dynamics of SFMs for anti-spoofing diverge sharply from ASR trends. While ASR continuously benefits from expanded multilingual coverage, deepfake detection hits a performance plateau at moderate scales. Specifically, XLS-R-128 consistently provides the optimal balance between in-domain performance and OOD generalization. Because spoofing detection relies primarily on low-level acoustic anomalies rather than semantic or linguistic modeling, moderate multilingual pretraining provides sufficient diversity to boost generalization without diluting target synthesis cues. Conversely, massive scaling introduces excessive language-specific noise that degrades performance. Crucially, multilingual pretraining also enhances robustness on English-only monolingual benchmarks, indicating that diverse linguistic exposure fosters broader representation invariance.

While larger parameter sizes offer complementary benefits in OOD settings, scale alone cannot offset poor pretraining data diversity; smaller, moderately multilingual models frequently outperform massive alternatives. Our analysis highlights the critical need for OOD evaluation, as saturation on standard benchmarks like 19LA masks vulnerabilities. In particular, diffusion- and flow-matching-based architectures, alongside cross-lingual distribution shifts, present severe failure modes for English-only monolingual SFMs. For performance, XLS-R-128 Large/XL yield the best robustness-to-scale trade-off, while mHuBERT-147 provides an efficient Base-scale alternative.

## 7 Conclusion

We conducted a comprehensive evaluation of 20 speech foundation models (SFMs) across 22 distinct evaluation conditions including vocoder artifacts, neural speech generators, multilingual shifts, and cross-accent variations. Our findings yield three key takeaways: first, multilingual pretraining establishes superior representation invariance over English-only SFMs; second, extreme multilingual scaling yields diminishing returns, with moderate coverage striking the best generalization trade-off; and third, model scale is secondary to pretraining diversity. Ultimately, XLS-R-128 emerges as the most stable frontend, proving that effective anti-spoofing depends heavily on learning acoustically robust, transferable speech representations.

## Limitations

Despite the breadth of our evaluation, several limitations remain. First, all models are fine-tuned on a single training benchmark (`ASVspoof 2019 LA`), which may introduce dataset-specific biases and limit the generality of some conclusions. Different trends may emerge under multi-dataset, multilingual, or continual training settings. Second, our experiments focus primarily on linear classification heads applied to SFM representations. More expressive downstream architectures may alter the relative importance of multilingual pretraining and model scale. Third, the analysis mainly relies on final-layer representations. Internal layer-wise behavior remain unexplored and may provide deeper insights into how SFMs encode synthesis artifacts across languages and spoofing methods.

Fourth, while our study covers a broad range of modern datasets, emerging generative paradigms continue to evolve rapidly. Future speech generation systems may exhibit substantially different artifact characteristics, potentially changing current robustness rankings. Finally, the observed multilingual scaling plateau is derived from a discrete set of publicly available models and should not be interpreted as a precise universal optimum. Additional studies with controlled pretraining corpora and scaling configurations are necessary to better understand the relationship between multilingual diversity, acoustic representation quality, and spoofing robustness.

## Ethical considerations

This work focuses on defensive applications of speech technology by improving the robustness and generalization of speech deepfake detection systems. All datasets, pretrained models, and evaluation protocols used in this study are publicly available research resources developed for scientific purposes. Robust spoofing detection is increasingly important due to the growing misuse potential of synthetic speech in fraud, impersonation, misinformation, and social attacks. By systematically analyzing the robustness limitations of current SFMs, our work aims to support the development of more reliable and secure detection systems.

At the same time, research on detection systems may indirectly provide insights into the weaknesses of current detectors and could potentially inform adversarial synthesis strategies. However, we do not release new spoofing generation systems, attack pipelines, or methods intended to circumvent detection. We believe that transparent evaluation and robust benchmarking are necessary for advancing trustworthy speech technology, and that the societal benefits of improving defensive deepfake detection outweigh the associated risks.

## References


Hashim Ali, Nithin Sai Adupa, Surya Subramani, and Hafiz Malik. 2026. A superb-style benchmark of self-supervised speech models for audio deepfake detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16512–16516.

Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2033–2044, New York, NY, USA. Association for Computing Machinery.

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. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J. Delp. 2025. Diffssd: A diffusion-based dataset for speech forensics. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Luis Buera, Héctor Delgado, Daniele Colibro, and Antonio Miguel. 2026. Combining ssl speech features, contextual transformers and mamba models for realistic audio spoofing detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16247–16251.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pages 3670–3674.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. 2022b. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6152–6156.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Interspeech 2021*, pages 2426–2430.

Phuong Tuan Dat and Tran Huy Dat. 2025. Xlsr-kanformer: A kan-intergrated model for synthetic speech detection. In *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, pages 1–6.

Phuong Tuan Dat, Duc-Tuan Truong, Long-Vu Hoang, and Nguyen Thi Thu Trang. 2026. Fine-grained frame modeling in multi-head self-attention for speech deepfake detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18082–18086.

Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Mathew Magimai.-Doss. 2026. Speech df arena: A leaderboard for speech deepfake detection models. *IEEE Open Journal of Signal Processing*, 7:73–81.

Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-Yi Lee, and Jyh-Shing Roger Jang. 2024. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 921–928.

Yassine El Kheir, Fabian Ritter-Guttierez, Arnab Das, Tim Polzehl, and Sebastian Möller. 2026. A parameter-efficient multi-scale convolutional adapter for synthetic speech detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18072–18076.

Yassine El Kheir, Younes Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. 2025. Comprehensive layer-wise analysis of SSL models for audio deepfake detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4070–4082, Albuquerque, New Mexico. Association for Computational Linguistics.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689.

Femi Folorunsho and Benedicta Frema Boamah. 2025. Deepfake technology and its impact: ethical considerations, societal disruptions, and security threats in ai-generated media. *International journal of information technology and management information systems*, 16(1):1060–1080.

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for deepfake speech detection. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Yu Guan, Wu Guo, Jie Zhang, and Zhijun Zhang. 2025. Fusing multi-layer features of the pre-trained model with grouped cross attention for spoofing speech detection. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 601–606.

Yunqi Hao, Minqiang Xu, Yihao Chen, Yanyan Liu, Liang He, Lei Fang, and Lin Liu. 2025. Integrating spectro-temporal cross aggregation and multi-scale dynamic learning for audio deepfake detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and sheng zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning*.

Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, and Shinji Watanabe. 2025. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 6:68–77.

Taewoo Kim, Guisik Kim, Choongsang Cho, and Young Han Lee. 2025. Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection. In *Interspeech 2025*, pages 5318–5322.

Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, and Mathew Maginai Doss. 2024. Exploring generalization to unseen audio data for spoofing: insights from SSL models. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 86–93.

Howard Lei and Eduardo Lopez. 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *Interspeech 2009*, pages 2323–2326.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. 2024. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 3585–3599, New York, NY, USA. Association for Computing Machinery.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and 1 others. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Douglas O'Shaughnessy. 2026. Review of automatic text-to-speech synthesis methods. *IEEE Transactions on Audio, Speech and Language Processing*, pages 1–23.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Tuan Dat Phuong, Long-Vu Hoang, and Huy Dat Tran. 2025. Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models. In *Interspeech 2025*, pages 5633–5637.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020a. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Interspeech 2020*, pages 4751–4755.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1).

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020b. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Jie Pu, Yuguang Yang, Ruirui Li, Oguz Elibol, and Jasha Droppo. 2021. Scaling Effect of Self-Supervised Speech Models. In *Interspeech 2021*, pages 1084–1088.

Ricardo Reimao and Vassilios Tzerpos. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10.

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. 2023. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech 2023*, pages 5281–5285.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Interspeech 2023*, pages 884–888.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. 2023. Ai-synthesized voice detection using neural vocoder artifacts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 904–912.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022a. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing.

In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022b. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 112–119.

Pablo Andrés Tamayo Flórez, Rubén Manrique, and Bernardo Pereira Nunes. 2023. Habla: A dataset of latin american spanish accents for voice anti-spoofing. In *Interspeech 2023*, pages 1963–1967.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech 2019*, pages 1008–1012.

Hoan My Tran, David Guennec, Philippe Martin, Aghilas Sini, Damien Lolive, Arnaud Delhay, and Pierre-François Marteau. 2024. Spoofed Speech Detection with a Focus on Speaker Embedding. In *Interspeech 2024*, pages 2080–2084.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. 2025a. Kan you hear the truth? audio deepfake detection with kolmogorov–arnold networks. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 561–565.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. 2025b. Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection. In *Interspeech 2025*, pages 5323–5327.

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. 2025c. Multi-level ssl feature gating for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11766–11775, New York, NY, USA. Association for Computing Machinery.

Hoan My Tran, Aghilas Sini, David Guennec, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. 2026. Kan we make models simpler for audio deepfake detection with kolmogorov–arnold networks? In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18057–18061.

Duc-Tuan Truong, Tianchi Liu, Junjie Li, Ruijie Tao, Kong Aik Lee, and Eng Siong Chng. 2025a. Addressing gradient misalignment in data-augmented training for robust speech deepfake detection. *arXiv preprint arXiv:2509.20682*.

Duc-Tuan Truong, Tianchi Liu, Junjie Li, Ruijie Tao, Kong Aik Lee, and Eng Siong Chng. 2026. Addressing gradient misalignment in data-augmented training for robust speech deepfake detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16252–16256.

Duc-Tuan Truong, Tianchi Liu, Ruijie Tao, Junjie Li, Kong Aik Lee, and Eng Siong Chng. 2025b. Qamo: Quality-aware multi-centroid one-class learning for speech deepfake detection. *arXiv preprint arXiv:2509.20679*.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, and 10 others. 2026. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825.

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. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pages 1194–1198.

Yang Xiao and Rohan Kumar Das. 2025. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 32:1276–1280.

Xi Xuan, Xuechen Liu, Wenxin Zhang, Yi-Cheng Lin, Xiaojian Lin, and Tomi Kinnunen. 2026. Wavespnet: Learnable wavelet-domain sparse prompt tuning for speech deepfake detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18047–18051.

Xi Xuan, Zimo Zhu, Wenxin Zhang, Yi-Cheng Lin, and Tomi Kinnunen. 2025. Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative. *arXiv preprint arXiv:2508.09294*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/~ idea/readings/rainbow. htm).*

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Mingru Yang, Yanmei Gu, Qianhua He, Yanxiong Li, Peirong Zhang, Yongqiang Chen, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. 2025a. Generalizable Audio Deepfake Detection via Hierarchical Structure Learning and Feature Whitening in Poincaré sphere. In *Interspeech 2025*, pages 2255–2259.

Mingru Yang, Yanmei Gu, Qianhua He, Peirong Zhang, Haolin He, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. 2025b. Generalizable audio deepfake detection via risk-aware style alignment and structural empirical risk minimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11600–11609, New York, NY, USA. Association for Computing Machinery.

Minjiao Yang, Kangfeng Zheng, Jujie Wang, Xiaoyu Zhang, and Yaru Zhao. 2026. Tri-attention fusion: Joint temporal-spectral and bidirectional modeling for speech spoofing detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18077–18081.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, and 1 others. 2023. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*, pages 3939–3943.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530.

Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. 2021. Fmfcc-a: a challenging mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*, pages 117–131. Springer.

Jinpeng Zhao, Jian Zhao, Yufei Zhou, Peijia Zheng, and Yusong Du. 2026. Robust deepfake audio detection via multi-level intermediate feature fusion. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18062–18066.

# A Speech foundation model details

We explore a wide suite of SFMs spanning diverse SSL, parameter dimensions, and pretraining language coverage to establish their transferability to spoofing generalization.

## A.1 `Wav2Vec2`

The `Wav2Vec2` architecture (Baevski et al., 2020) is a multi-layer Transformer speech encoder optimized via a masked latent contrastive loss. The network constructs contextualized representation matrices directly from raw sound waves by predicting quantized codebook vectors at masked temporal locations. We evaluate English-centric variants alongside `Wav2Vec2 Vox.`, which is trained on the multilingual `VoxPopuli` corpus (Wang et al., 2021), establishing an empirical base for exploring linguistic scope.

## A.2 `HuBERT`

The Hidden-Unit BERT (`HuBERT`) framework (Hsu et al., 2021) replaces contrastive objectives with a masked prediction task targeting discrete cluster indices discovered via offline $k$-means steps.

This acoustic assignment paradigm encourages the model to infer frame-level phonemic structures without manual transcriptions. We evaluate the standard English HuBERT along with mHuBERT-147 (Zanon Boito et al., 2024), a compact multilingual representation model pre-trained on 147 languages to observe cross-lingual spoofing robustness.

### A.3 UniSpeech-SAT

The UniSpeech-SAT architecture (Chen et al., 2022b) combines standard self-supervised speech tasks with multi-speaker contrastive learning to enhance speaker discrimination while preserving underlying phonemic sequences. We analyze the standard UniSpeechSAT network alongside UniSpeechSAT+, a configuration trained on more acoustically diverse English audio data. Their speaker-entangled hidden spaces are highly applicable to deepfake classification, as preserving speaker identity cues can help isolate subtle processing artifacts.

### A.4 WavLM

The WavLM framework (Chen et al., 2022a) expands on clustering-based masked token prediction by introducing a gated relative position bias mechanism and an innovative utterance mixing pretraining strategy. This formulation enables the model to simultaneously extract acoustic content and speaker characteristics. We examine both the standard WavLM and WavLM+, with the latter optimized over an augmented corpus to yield higher representation fidelity. These systems are expected to show robust resistance against channel noise and speaker drift.

### A.5 XLS-R

The XLS-R model family (Babu et al., 2022) scales the cross-lingual Wav2Vec2 structure to massively multilingual speech representations by training on audio data from over 100 distinct languages. We examine XLS-R-53 (Conneau et al., 2021) and XLS-R-128, which vary primarily in parameter allocation and total pre-training data volume. Their extensive multi-language exposure is designed to yield robust representation transfer across heterogeneous dialects and recording mediums, making them strong candidates for out-of-domain evaluation.

### A.6 MMS

The Massively Multilingual Speech (MMS) architecture (Pratap et al., 2024) is a large-scale frontend trained on over one thousand distinct languages utilizing self-supervised loss formulations derived from Wav2Vec2. Its unmatched linguistic coverage allows the model to learn deep universal acoustic concepts, potentially minimizing performance degradation when encountering unseen generation algorithms or cross-lingual target distributions.

### A.7 Omnilingual

The Omnilingual network (Omnilingual et al., 2025) is an advanced speech foundation frontend designed to produce unified language representations across a vast array of global domains. Utilizing massively scale multi-language optimization regimes, it captures language-agnostic speech properties that generalize across downstream tasks. This architecture allows us to assess whether universal linguistic abstractions improve deepfake classification accuracy under extreme distribution shifts.

**Model scale and language scope.** To analyze the role of model capacity, we group SFMs into Base, Large, and XL categories according to their parameter count. We also distinguish between English-only and multilingual pretraining regimes. English-only models are trained primarily on English corpora and may learn more specialized acoustic representations, whereas multilingual models are exposed to broader phonetic and acoustic diversity, potentially improving robustness to out-of-domain and multilingual spoofing scenarios. This model suite enables a systematic comparison of how SSL objectives, parameter scale, and multilingual exposure affect the transferability of SSL representations for speech deepfake detection.

## B Dataset details

In this section, we provide exhaustive descriptions of the benchmark databases utilized in this study (Tab. 4) to evaluate spoofing countermeasures (CMs).

### B.1 ASVspoof 2019

The ASVspoof 2019 logical access (19LA[1]) scenario (Wang et al., 2020) serves as a foundational benchmark for assessing CMs against synthetic speech attacks directed at ASV systems. The 19LA

[1] https://doi.org/10.7488/ds/2555

track targets logical access threats, where spoofed audio is synthesized via TTS and VC systems. It integrates bona fide speech from the VCTK corpus (Yamagishi et al., 2019) with spoofed instances created using 17 distinct generative mechanisms (6 known techniques in the training and development partitions, and 13 unknown techniques in the evaluation set). This configuration drives the development of robust classifiers capable of identifying modern speech synthesis variations under controlled acoustic environments.

### B.2 ASVspoof 2021

The ASVspoof 2021 (Liu et al., 2023) logical access (21LA[2]) framework builds on the 19LA partition by subjecting the evaluation samples to realistic network transmission effects, including various telephony and VoIP lossy codecs, to mimic genuine communication channels. In parallel, the ASVspoof 2021 deepfake (21DF[3]) track broadens the detection objective beyond ASV vulnerabilities to generic audio deepfake identification. This dataset compiles authentic recordings alongside spoofed utterances processed through diverse lossy compression algorithms and bitrates, introducing channel artifacts to challenge detector robustness.

### B.3 ASVspoof 5

The ASVspoof 5 (Wang et al., 2026) (25DF[4]) challenge introduces a highly intricate database for evaluating anti-spoofing and deepfake detection pipelines within realistic operational settings. Moving away from the pristine, studio-recorded environments of earlier challenge iterations, ASVspoof 5 leverages crowdsourced authentic speech spanning roughly 2,000 distinct speakers under heterogeneous acoustic backgrounds, drawn from the English portion of the Multilingual Librispeech (MLS) framework (Pratap et al., 2020b). The spoofed utterances are generated using advanced neural TTS/VC models and novel adversarial algorithms, with the evaluation sets featuring unseen codecs and acoustic compression to encourage real-world, in-the-wild generalization.

### B.4 SpoofCeleb

The SpoofCeleb (SC[5]) benchmark (Jung et al., 2025) provides a large-scale framework for voice deepfake detection, compiling real-world, in-the-wild authentic audio from 1,251 unique public figures derived from the VoxCeleb1 corpus via an automated processing chain encompassing transcription, diarization, noise attenuation, and quality thresholding. The corpus contains over 2.5 million files, where the spoofed counterparts are synthesized through 23 diverse TTS systems. The evaluation partition contains 40 speakers and 9 attack methodologies that emphasize high environmental reverberation, background noise, and wide speaker diversity to accurately mirror practical deepfake threats.

### B.5 In-the-Wild

The In-the-Wild (ITW[6]) database (Müller et al., 2022) features roughly 37.9 hours of audio fragments from 58 English-speaking public figures and politicians. It contains 20.7 hours of authentic speech extracted from open-source podcasts and public addresses, paired with 17.2 hours of forged speech collected from 219 unconstrained internet sources. The audio clips are systematically balanced according to speech style, emotional delivery, environmental background noise, and length to represent real-world deployment. The ITW dataset contains unseen voice synthesis architectures, social media encoding artifacts, and varied transmission acoustics, acting as a rigorous OOD evaluation where models trained on standard datasets frequently suffer dramatic performance degradation.

### B.6 Fake-or-Real

The Fake-or-Real (FoR[7]) corpus (Reimao and Tzerpos, 2019) comprises more than 195,000 English utterances, containing approximately 87,000 synthetic and 111,000 authentic audio files created via prominent industrial and research TTS architectures. The genuine recordings are collected from heterogeneous open-source datasets and real-world media, incorporating vast diversity in speaker demographics, physiological age, regional accents, microphone hardware, and room impulse responses. The database is distributed in multiple variations, including FOR-original (untruncated utterances), FOR-norm (loudness normalized and balanced), FOR-2seconds (fixed 2-second windows), and FOR-rerecorded (re-played and recorded to capture realistic channel playback distortions).

[2] https://doi.org/10.5281/zenodo.4817650
[3] https://doi.org/10.5281/zenodo.4835107
[4] https://doi.org/10.5281/zenodo.14498691
[5] https://huggingface.co/datasets/jungjee/spoofceleb
[6] https://deepfake-total.com/in_the_wild
[7] https://bil.eecs.yorku.ca/datasets

This study leverages the FOR-original variant to analyze OOD transferability.

### B.7 MLAAD

The Multi-Language Audio Anti-Spoofing Dataset (Müller et al., 2024) (MLAAD[8]) is an expansive multilingual evaluation suite engineered to combat the linguistic bias found in traditional benchmarks and analyze countermeasure generalization across wide cross-lingual contexts. It provides over 570 hours of synthetic speech ( 243,000 utterances) generated via 119 diverse TTS pipelines. The underlying authentic speech is drawn from the M-AILABS collection in 8 native languages—English (MLAAD-EN), French (M-FR), German (M-DE), Italian (M-IT), Polish (M-PL), Russian (M-RU), Spanish (M-ES), and Ukrainian (M-UK)—and expands to 40 additional target languages via advanced machine translation. Audio samples are delivered at a 22,050 Hz sampling rate, with evaluation trials featuring synthetic noise injection, background music, and lossy codec compression.

### B.8 Deepfake Cross-Lingual Evaluation

The Deepfake Cross-Lingual Evaluation (DC[9]) corpus (Ba et al., 2023) is a specialized bilingual dataset structured to quantify the cross-lingual transfer capabilities of voice spoofing detectors by highlighting structural failure modes under cross-language distribution shifts. It consists of two discrete subsets: DECRO-E (English) and DC-C (Chinese). The genuine recordings within the Chinese partition are sourced from the FMFCC-A database (Zhang et al., 2021), whereas the corresponding English files are generated using commercial web-based TTS APIs. Spoofed samples are constructed using various advanced TTS and VC platforms, emphasizing cross-lingual evaluation settings where networks are trained on one language and tested on another.

### B.9 LibriSeVoc

The LibriSeVoc (LSV[10]) benchmark (Sun et al., 2023) is an acoustic database engineered to assess voice deepfake systems on the footprints left by neural vocoders rather than focusing on end-to-end TTS/VC pipelines. Built upon the clean, multi-speaker LibriTTS corpus (Zen et al., 2019), it provides 92,407 audio files spanning roughly 244 hours at a 24 kHz sampling rate. The corpus includes 13,201 authentic recordings and 79,206 spoofed instances generated through self-vocoding mechanisms, where original speech is re-synthesized using six distinct neural vocoder paradigms across autoregressive, GAN-based, and diffusion-based architectures.

### B.10 DFADD

The Diffusion and Flow-Matching Based Audio Deepfake Dataset (Du et al., 2024) (DFADD[11]) targets highly natural synthetic speech produced by modern diffusion and flow-matching TTS generators. Constructed using source files from the VCTK database, it features approximately 180 hours of total speech data, comprised of 44,455 authentic utterances and 163,500 spoofed audio instances, posing an empirical challenge to conventional classifiers.

### B.11 DIFFSSD

The Diffusion-Based Synthetic Speech Dataset (DIFFSSD[12]) (Bhagtani et al., 2025) addresses generalization errors in countermeasures when subjected to high-fidelity generations from contemporary diffusion-based speech models. It contains approximately 200 hours of audio data, where real utterances are gathered from the LJ Speech (Ito and Johnson, 2017) and LibriSpeech (Panayotov et al., 2015) databases across 74 unique speakers. These are paired with 70,000 synthetic audio clips generated via 10 modern text-to-speech architectures, incorporating 8 open-source diffusion pipelines and 2 commercial platforms.

### B.12 Audio Deepfake Detection 2023

The Audio Deepfake Detection challenge (Yi et al., 2023) (ADD[13] Track 1.2) evaluates countermeasures under an adversarial "fake game" environment, where synthetic attacks are explicitly optimized to escape detection. The evaluation partition contains two sequential steps: Round 1 (ADD-R1) containing 111,976 test utterances (80,000 real and 31,976 fake files created via TTS/VC and partial challenge submissions), and Round 2 (ADD-R2) with 118,477 test recordings (87,500 real and 30,977

[8] https://deepfake-total.com/mlaad
[9] https://doi.org/10.5281/zenodo.7601506
[10] https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts
[11] https://github.com/isjwdu/DFADD
[12] https://huggingface.co/datasets/purdueviperlab/diffssd
[13] http://addchallenge.cn/databases2023

spoofed instances integrating more aggressive adversarial optimization). Unlike typical English-focused datasets, ADD Track 1.2 centers on large-scale Mandarin speech with participant-generated deepfakes to emulate real-world security threats.

### B.13 HABLA

The HABLA[14] dataset (Tamayo Flórez et al., 2023) stands as the premier public speech anti-spoofing corpus dedicated to Spanish, highlighting regional Latin American accent profiles (Argentinian, Chilean, Colombian, Peruvian, and Venezuelan) across 162 male and female speakers. It gathers 22,816 bona fide samples and approximately 58,000 spoofed records generated through six modern speech synthesis frameworks, including cross-accent voice conversion. This collection highlights dialectal diversity to test the generalization limits of CM architectures under fine-grained regional phonemic drift.

## C Evaluation metric

The EER corresponds to the operating point at which the false acceptance (FA) rate equals the false rejection (FR) rate. In the context of spoofing CMs, these rates are functions of a decision threshold $\tau_{\text{CM}}$ applied to the detection score:

$$P_{\text{fa}}^{\text{CM}}(\tau_{\text{CM}}) = \frac{\#\{\text{spoofed with score} > \tau_{\text{CM}}\}}{\#\{\text{spoofed trials}\}}, \tag{1}$$

$$P_{\text{miss}}^{\text{CM}}(\tau_{\text{CM}}) = \frac{\#\{\text{bona fide with score} \leq \tau_{\text{CM}}\}}{\#\{\text{bona fide trials}\}}. \tag{2}$$

The EER is obtained by sweeping $\tau_{\text{CM}}$ to the operating point where $P_{\text{fa}}^{\text{CM}} = P_{\text{miss}}^{\text{CM}}$. Lower EER values indicate stronger spoofing detection performance. In our framework, detection scores are computed as log-likelihood ratios (LLR):

$$\text{LLR}_t = \log p(X_t \mid \mathcal{H}_0) - \log p(X_t \mid \mathcal{H}_1), \tag{3}$$

where $X_t$ denotes the input utterance, $\mathcal{H}_0$ the bona fide hypothesis, and $\mathcal{H}_1$ the spoofed hypothesis.

## D Implementation setup

As shown in Table 3, all downstream models are implemented using PyTorch and fine-tuned using pretrained SFMs sourced from HuggingFace. Due to the variable durations of the input waveforms, dynamic batching is employed, where audio sequences within each batch are padded to match the maximum sample length of that specific batch.

To counteract the severe class imbalance inherent in the 19LA training partition, we utilize a weighted cross-entropy loss function. This objective assigns a static weight of 0.9 to the minority bona fide (real) class and 0.1 to the majority spoofed (fake) class, ensuring that the gradient updates remain sensitive to genuine speech distribution patterns. To improve OOD generalization and simulate real-world channel variants, we also apply RawBoost data augmentation (Tak et al., 2022a)[15] to add various types of noise including linear and nonlinear convolutive, impulsive signal-dependent additive, stationary additive, and randomly colored noises during training.

Optimization is performed using the Adam optimizer with a fixed learning rate of $2.5 \times 10^{-6}$ and a weight decay coefficient of $1.0 \times 10^{-4}$ to prevent overfitting. Every SFM is fine-tuned for a maximum of 20 epochs using a batch size of 5 samples on a NVIDIA A100 GPU. To optimize training efficiency, we implement an early-stopping mechanism with a patience threshold of 3 epochs, monitoring the validation loss to determine final convergence.

| Configurations | Setting |
|---|---|
| Batch size | 5 |
| Epochs | 20 |
| GPU | 1 NVIDIA A100 |
| Optimizer | Adam |
| Learning rate | $2.5 \times 10^{-6}$ |
| Weight decay | $1e^{-4}$ |
| Weighted cross-entropy loss | 0.9 for real, 0.1 for fake |
| Early-stop patience | 3 |
| Data augmentation | RawBoost |

Table 3: Hyperparameters details of the models.

[14] https://doi.org/10.5281/zenodo.7370804

[15] https://github.com/TakHemlata/RawBoost-antispoofing

| Dataset | | Language | # Bona fide | # Spoofed | Attack |
|---|---|---|---|---|---|
| **Training** | | | | | |
| ASVspoof 2019 logical access (19LA) (Wang et al., 2020) | | English | 2,580 | 22,800 | TTS, VC |
| **Development** | | | | | |
| ASVspoof 2019 logical access (19LA) (Wang et al., 2020) | | English | 2,548 | 22,296 | TTS, VC |
| **Evaluation** | | | | | |
| ASVspoof 2019 logical access (19LA) (Wang et al., 2020) | | English | 7,355 | 63,882 | TTS, VC |
| ASVspoof 2021 logical access (21LA) deepfake (21DF) (Liu et al., 2023) | | English | 14,816<br>14,869 | 133,360<br>519,059 | TTS, VC |
| ASVspoof5 (25DF) (Wang et al., 2026) | | English | 138,688 | 542,086 | TTS, VC, AT |
| Fake-or-Real (FoR) (Reimao and Tzerpos, 2019) | | English | 2,264 | 2,370 | TTS |
| In-the-Wild (ITW) (Müller et al., 2022) | | English | 19,963 | 11,816 | Unknown |
| Diffusion-based Synthetic Speech Dataset (DFADD) (Du et al., 2024) | | English | 755 | 3,000 | TTS |
| Librisevoc (LSV) (Sun et al., 2023) | | English | 2,641 | 15,846 | Vocoded |
| SpoofCeleb (SC) (Jung et al., 2025) | | English | 9,113 | 82,017 | TTS |
| Deepfake Cross-Lingual Evaluation (DECRO) | DECRCO-EN<br>D-CH (Ba et al., 2023) | English<br>Chinese | 4,306<br>6,109 | 14,884<br>12,015 | TTS, VC |
| Audio Deepfake Detection (ADD-R2) (Yi et al., 2023) | | Chinese | 87,500 | | |
| HABLA (Tamayo Flórez et al., 2023) | | Spanish | 9,057 | 23,270 | TTS, VC |
| Multi-Language Audio Anti-Spoofing Dataset (MLAAD) | MLAAD-EN | English | 28,233 | 36,000 | TTS |
| | M-FR | French | 7,686 | 8,000 | |
| | M-DE | German | 8,696 | 9,000 | |
| | M-ES (Müller et al., 2024) | Spanish | 6,655 | 7,000 | |
| | M-IT | Italian | 7,611 | 8,000 | |
| | M-PL | Polish | 5,489 | 6,000 | |
| | M-UK | Ukrainian | 4,709 | 5,000 | |
| | M-RU | Russian | 4,540 | 5,000 | |
| Diffusion-based Synthetic Speech Dataset (DIFFSSD) (Bhagtani et al., 2025) | | English | 12,113 | 42,500 | TTS |

AT: Adversarial attack.

Table 4: Statistics of datasets used in the study.

| Size | Model | Lang. | In-Domain | | | Voc. | In-the-Wild | | | Diff/Flow | | Other OOD | | | Avg. | Pool. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 19LA | 21LA | 21DF | LSV | ITW | FoR | SC | DFADD | DIFSSD | MLAAD-EN | DECRO-EN | 25DF | | |
| Base | Wav2Vec 2.0 | EN | 1.33 | 5.99 | 6.81 | 7.84 | 14.55 | 14.52 | 33.03 | 19.87 | 31.47 | 22.59 | 6.39 | 20.42 | 15.40 | 20.38 |
| | HuBERT | EN | 1.23 | 5.96 | 7.56 | 10.80 | 15.55 | 4.14 | 40.30 | 11.10 | 26.61 | 27.98 | 8.18 | 28.20 | 15.63 | 29.16 |
| | UniSpeechSAT | EN | 1.14 | 3.83 | 6.46 | 9.59 | 17.04 | 5.07 | 27.77 | 16.31 | 29.30 | 22.92 | 7.59 | 20.32 | 13.95 | 20.39 |
| | UniSpeechSAT+ | EN | 0.37 | 1.92 | 6.40 | 7.65 | 12.55 | 9.41 | 28.26 | 14.97 | 23.23 | 15.57 | 3.22 | 14.34 | 11.49 | 14.47 |
| | WavLM | EN | 2.52 | 4.60 | 8.51 | 11.24 | 11.82 | 4.42 | 24.70 | 17.36 | 48.05 | 26.02 | 7.81 | 18.90 | 15.50 | 21.67 |
| | WavLM+ | EN | 0.52 | 2.04 | 5.21 | 3.10 | 11.62 | 10.73 | 23.72 | 14.15 | 45.68 | 19.32 | 4.65 | 16.57 | 13.11 | 20.32 |
| | Wav2Vec2 Vox. | Multi | 0.36 | 3.37 | 9.09 | 13.59 | 11.36 | 12.82 | 30.10 | 22.11 | 27.22 | 18.11 | 4.37 | 22.07 | 14.55 | 19.74 |
| | mHuBERT-147 | Multi | 0.44 | 1.41 | 4.34 | 2.61 | 10.84 | 2.20 | 20.70 | 13.77 | 19.67 | 18.38 | 3.69 | 15.53 | 9.46 | 16.82 |
| Large | Wav2Vec 2.0 | EN | 0.85 | 2.48 | 10.60 | 29.38 | 14.28 | 7.77 | 35.60 | 14.58 | 28.46 | 21.13 | 4.74 | 19.99 | 15.82 | 18.97 |
| | HuBERT | EN | 0.65 | 3.57 | 5.95 | 7.42 | 14.55 | 13.08 | 37.22 | 10.07 | 21.34 | 7.47 | 3.30 | 16.83 | 11.79 | 16.02 |
| | UniSpeechSAT | EN | 0.85 | 2.91 | 3.58 | 6.33 | 10.80 | 55.07 | 35.71 | 13.27 | 45.22 | 64.91 | 8.96 | 7.07 | 22.06 | 17.95 |
| | WavLM | EN | 0.36 | 4.24 | 3.65 | 9.06 | 10.94 | 5.78 | 54.33 | 12.06 | 52.29 | 50.41 | 2.11 | 73.01 | 23.19 | 39.37 |
| | Wav2Vec2 Vox. | Multi | 0.67 | 5.48 | 8.70 | 20.26 | 14.91 | 24.38 | 39.34 | 15.23 | 37.24 | 37.82 | 4.20 | 22.76 | 19.25 | 21.01 |
| | XLS-R-53 | Multi | 0.22 | 2.44 | 3.86 | 7.62 | 10.62 | 6.24 | 31.79 | 19.87 | 19.50 | 13.22 | 5.04 | 18.63 | 11.59 | 16.28 |
| | XLS-R-128 | Multi | 0.10 | 2.36 | 1.46 | 1.75 | 4.69 | 0.93 | 19.57 | 17.51 | 10.16 | 12.44 | 3.48 | 13.47 | 7.34 | 11.90 |
| | MMS | Multi | 0.08 | 3.06 | 2.54 | 2.27 | 7.08 | 1.68 | 20.50 | 8.59 | 12.06 | 12.31 | 2.27 | 13.96 | 7.20 | 13.31 |
| | Omnilingual | Multi | 0.16 | 1.31 | 3.13 | 6.11 | 5.96 | 0.48 | 20.22 | 16.54 | 16.80 | 15.34 | 2.81 | 12.13 | 8.42 | 11.89 |
| XL | XLS-R-128 | Multi | 0.14 | 1.38 | 1.68 | 1.98 | 5.73 | 1.73 | 15.69 | 4.63 | 10.58 | 6.72 | 1.42 | 12.62 | 5.36 | 12.37 |
| | MMS | Multi | 0.21 | 1.08 | 2.08 | 3.34 | 4.92 | 1.14 | 18.20 | 6.23 | 11.31 | 12.65 | 2.30 | 10.24 | 6.14 | 11.31 |
| | Omnilingual | Multi | 0.10 | 1.16 | 2.63 | 3.03 | 4.22 | 1.23 | 15.09 | 9.80 | 12.89 | 9.66 | 2.63 | 12.45 | 6.24 | 10.61 |

Table 5: Performance (EER %, ↓) of speech foundation models across evaluated English benchmarks. Models are grouped by scale (Base/Large/XL) and pretraining scope (**EN** = English-only, **Multi** = multilingual). **Avg.** is the mean EER across all 12 benchmarks; **Pool.** is the pooled EER.

| Size | Model | Lang. | ADD-R2 | D-CH | HABLA | M-DE | M-ES | M-FR | M-IT | M-PL | M-RU | M-UK | Avg. | Pool. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BASE | Wav2Vec 2.0 | EN | 34.64 | 26.31 | 17.10 | 43.91 | 40.80 | 29.22 | 22.20 | 33.32 | 26.66 | 36.50 | 31.07 | 26.04 |
| | HuBERT | EN | 34.12 | 19.40 | 19.28 | 42.48 | 32.44 | 25.89 | 21.54 | 37.11 | 26.54 | 43.98 | 30.28 | 30.62 |
| | UniSpeechSAT | EN | 32.80 | 20.68 | 16.17 | 44.10 | 36.48 | 27.19 | 20.20 | 30.40 | 23.20 | 34.18 | 28.54 | 24.30 |
| | UniSpeechSAT+ | EN | 29.20 | 20.79 | 7.87 | 33.34 | 28.77 | 23.14 | 21.21 | 26.90 | 17.90 | 26.72 | 23.58 | 18.30 |
| | WavLM | EN | 30.68 | 20.75 | 19.61 | 48.21 | 47.36 | 48.09 | 26.99 | 46.67 | 30.77 | 38.42 | 35.75 | 28.43 |
| | WavLM+ | EN | 33.91 | 17.74 | 7.73 | 44.99 | 39.14 | 30.16 | 21.76 | 34.74 | 24.06 | 38.46 | 29.27 | 24.87 |
| | Wav2Vec2 Vox. | Multi | 30.26 | 20.56 | 11.22 | 29.28 | 22.28 | 21.91 | 14.87 | 16.73 | 20.90 | 28.52 | 21.65 | 20.98 |
| | mHuBERT-147 | Multi | 34.93 | 10.16 | 12.03 | 40.05 | 24.00 | 18.84 | 12.22 | 14.34 | 20.16 | 29.02 | 21.58 | 18.06 |
| LARGE | Wav2Vec 2.0 | EN | 23.96 | 17.83 | 12.58 | 46.85 | 28.79 | 31.34 | 24.55 | 40.37 | 25.44 | 27.28 | 27.90 | 21.82 |
| | HuBERT | EN | 24.63 | 21.98 | 7.12 | 27.88 | 14.86 | 17.90 | 11.22 | 24.78 | 13.34 | 20.64 | 18.44 | 20.44 |
| | UniSpeechSAT | EN | 26.40 | 22.57 | 7.19 | 57.22 | 60.60 | 51.57 | 66.30 | 67.43 | 54.94 | 53.32 | 46.75 | 22.48 |
| | WavLM | EN | 52.54 | 19.33 | 5.62 | 46.63 | 47.37 | 42.55 | 43.65 | 46.18 | 68.62 | 59.36 | 43.19 | 40.27 |
| | Wav2Vec2 Vox. | Multi | 29.60 | 27.04 | 4.80 | 56.22 | 53.45 | 50.36 | 60.79 | 57.72 | 57.58 | 64.98 | 46.25 | 24.49 |
| | XLS-R-53 | Multi | 20.73 | 20.66 | 5.71 | 24.61 | 13.31 | 16.35 | 12.75 | 16.82 | 14.26 | 20.50 | 16.57 | 26.29 |
| | XLS-R-128 | Multi | 22.78 | 13.08 | 2.25 | 16.19 | 8.87 | 6.39 | 5.62 | 8.22 | 7.82 | 11.24 | 10.24 | 21.96 |
| | MMS | Multi | 26.90 | 6.37 | 3.01 | 28.20 | 12.37 | 15.30 | 6.74 | 10.85 | 10.28 | 15.40 | 13.54 | 25.43 |
| | Omnilingual | Multi | 24.81 | 11.76 | 4.71 | 25.46 | 17.87 | 15.56 | 11.65 | 14.34 | 13.06 | 16.60 | 15.58 | 13.45 |
| XL | XLS-R-128 | Multi | 24.87 | 10.23 | 1.32 | 22.03 | 11.63 | 9.47 | 5.80 | 7.56 | 7.24 | 10.90 | 11.11 | 13.74 |
| | MMS | Multi | 27.01 | 11.10 | 2.52 | 29.46 | 14.84 | 15.34 | 7.75 | 11.90 | 11.74 | 11.92 | 14.36 | 12.73 |
| | Omnilingual | Multi | 24.56 | 10.10 | 2.16 | 25.29 | 12.70 | 10.53 | 6.48 | 9.71 | 8.54 | 11.12 | 12.12 | 13.37 |

Table 6: Performance (EER %, ↓) of speech foundation models across evaluated non-English benchmarks. Models are grouped by scale (BASE/LARGE/XL) and pretraining scope (**EN** = English-only, **Multi** = multilingual). **Avg.** is the mean EER across all 10 benchmarks; **Pool.** is the pooled EER.