Robust Multilingual Audio Deepfake Detection Through Hybrid Modeling

Candy Olivia Mawalim
Japan Advanced Institute of Science and Technology
Nomi, Japan
candylim@jaist.ac.jp

Yutong Wang
Japan Advanced Institute of Science and Technology
Nomi, Japan
s2310404@jaist.ac.jp

Aulia Adila
Japan Advanced Institute of Science and Technology
Nomi, Japan
adila@jaist.ac.jp

Shogo Okada
Japan Advanced Institute of Science and Technology
Nomi, Japan
okada-s@jaist.ac.jp

Masashi Unoki
Japan Advanced Institute of Science and Technology
Nomi, Japan
unoki@jaist.ac.jp

Abstract

The increasing sophistication of AI-generated human voice poses a significant threat, demanding robust detection systems that can generalize effectively across diverse linguistic environments and synthesis techniques. In response to the SAFE Challenge, this paper introduces a novel approach to multilingual audio deepfake detection. Our primary contribution lies in the comprehensive study of deepfake detection using a multilingual speech corpus encompassing 17 languages and a broad spectrum of synthesis methods and acoustic conditions, designed to enable more realistic and challenging evaluations. To optimally utilize this diverse data, we propose a hybrid detection model that synergistically combines the strengths of end-to-end RawNet and AASIST architectures with language-agnostic representations learned from a multilingual self-supervised learning model. Additionally, we explore the efficacy of RawBoost data augmentation in enhancing robustness against real-world noise. Our experimental evaluation demonstrates promising generalization in generated audio detection, achieving approximately 73% balanced accuracy across multilingual data and unseen synthesis algorithms.

CCS Concepts

• **Security and privacy** → **Biometrics**; • **Applied computing** → Computer forensics; • **Computing methodologies** → **Language resources**;

Keywords

multilingual, speech synthesis, dataset, generated voice, deepfake

IH&MMSEC '25, San Jose, CA, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1887-8/25/06
https://doi.org/10.1145/3733102.3736706

ACM Reference Format:
Candy Olivia Mawalim, Yutong Wang, Aulia Adila, Shogo Okada, and Masashi Unoki. 2025. Robust Multilingual Audio Deepfake Detection Through Hybrid Modeling. In *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSEC '25), June 18–20, 2025, San Jose, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3733102.3736706

1 Introduction

Recent advancements in deep learning techniques have significantly enhanced the ability to synthesize highly realistic human speech, enabling the widespread creation of machine-generated voices [40, 45]. While the innovations in machine-generated voices have fueled numerous positive applications in accessibility, entertainment, and virtual communication, they have also introduced potential threats, such as deepfake attacks, which pose serious risks to security and trust in digital communications.

A key requirement for machine-generated voice detection systems is generalizability [8, 25]. Creating strong detection systems that can reliably identify generated voices across different languages and the ever-changing landscape of synthesis techniques is challenging. Linguistic diversity, rapid evolution of synthesis technologies, dataset limitations, evaluation protocols, model robustness, and computational efficiency have become open challenges in this research [7]. Synthetic speech, for example, has reached a level where it can be nearly indistinguishable from human speech (pristine) in terms of intelligibility and naturalness, exposing the limitations of traditional evaluation and detection metrics, e.g., Mean Opinion Score (MOS) [28]. Hence, it is crucial to have a method to detect and localize differences between pristine and generated outputs, as current systems may fail to identify high-quality synthetic content. Furthermore, the lack of comprehensive, diverse, and standardized datasets that cover the full range of languages and synthesis techniques also limits the ability of detection systems to generalize. Achieving generalizability across multiple languages is critical to ensure inclusive protection beyond high-resource languages, as the development of speech generation technologies has expanded into more multilingual settings [4, 18, 30].

To address these emerging challenges, the SAFE (Synthetic Audio Forensics Evaluation)[1] Challenge has been launched to drive

[1] https://stresearch.github.io/SAFE/

innovation in the detection and attribute of synthetic and manipulated audio artifacts. Recognizing the challenge's consideration of realistic forensic scenarios (which may include various data quality and diverse processing [10]), this work directly addresses SAFE by presenting a novel approach for robust detection models. Our aim is to effectively distinguish between human and machine-generated speech across multilingual and multisource datasets. To achieve this, we utilized a comprehensive multilingual speech corpus from multiple open-source resources, namely JMAD dataset, subjecting them to rigorous curation to ensure both reliability and broad representativeness. JMAD dataset spans 17 languages and encompasses a wide spectrum of synthesis methods, recording conditions, and speaker demographics, explicitly designed to enhance model generalization beyond existing datasets.

Building on the foundation of our comprehensive multilingual dataset and the goal of robust detection, we implemented a hybrid modeling strategy. This approach strategically integrates the strengths of both end-to-end and self-supervised learning-based models to enhance the system's ability to distinguish differences between human- and machine-generated speech. To thoroughly evaluate the effectiveness and limitations of existing detection methods, we performed a detailed analysis of the model generalization. This investigation specifically focused on performance variations across the diverse languages and spoofing algorithms represented in our dataset, aiming to uncover inherent strengths and potential vulnerabilities in current approaches when faced with such varied data. Our analysis provides critical insights into the challenges of cross-lingual generalization in the context of generated speech detection.

Our key contributions significantly advance the field of machine-generated voiced detection by:

- Conducting a more comprehensive evaluation using a large-scale multilingual and multi-source corpus;
- Proposing a hybrid model leveraging state-of-the-art methods for cross-lingual and cross-source robustness;
- Analyzing the generalization challenges of detecting generated, processed, and laundered speech on completely unseen data.

2 Toward Multilingual Generated Voice Detection

Early research in generated voice detection or audio deepfake detection (ADD) was predominantly focused on English or a limited set of languages [19, 21, 45]. For instance, the well-known ASVspoof[2] challenge series served as a key benchmark in English [9, 38, 39, 41, 42]. While ADD[3] challenges have also been conducted in Chinese [43, 44], these benchmarks have largely driven the development and evaluation of detection systems using genuine human (pristine) voices often recorded under controlled laboratory conditions. Recognizing the limitations of this monolingual focus, recent efforts have begun to address the critical challenges of multilingualism. This section outlines the major existing work that historically tackles the complexities of detecting human and machine-generated speech across multiple languages.

2.1 Existing Datasets for Generated Voice Detection

Open-source datasets comprising both human and machine-generated speech play a pivotal role in fostering the development of detection systems. One of the most widely adopted benchmark datasets is provided by the ASVspoof Challenge [9, 20, 38, 39, 41]. Its latest edition is built on the MLS English dataset [31] and features stronger attacks, including advanced text-to-speech (TTS) and voice conversion (VC) algorithms designed to fool automatic speaker verification (ASV) and countermeasure (CM) subsystems, as well as adversarial attacks (AT). Codecs are also applied to both human and machine-generated speech to simulate realistic audio transmission conditions.

Similarly, the ADD Challenge [43, 44], which aims to address more complex real-life scenarios, has released datasets based on publicly available Chinese Mandarin corpora—AISHELL-1 [5], AISHELL-3 [34], and AISHELL-4 [11]. To generate fake audio samples in ADD datasets, researchers utilized a combination of advanced synthesis, character-level mixing, noise addition, and audio transcoding. These generation include both traditional and advanced neural network-based speech synthesis and voice conversion technologies, ensuring a wide variety of fake samples for robust detection research.

Other studies investigating cross-lingual performance in audio deepfake detection have indicated that a mismatch between the languages used for training and testing leads to a degradation in detection performance [1, 3]. To address this issue, Ba et al. constructed a novel cross-lingual evaluation dataset known as the DECRO (Deepfake Cross-Lingual) benchmark, which features spoofed speech in the two most widely spoken languages globally: English and Chinese [3]. Their findings suggest that models trained on English deepfake data can transfer knowledge of spoofing artifacts to other languages, particularly when domain adaptation techniques are employed to mitigate language dependency.

Subsequently, several datasets have been developed to support languages beyond English and Chinese. A notable multilingual datasets is MLAAD (The Multi-Language Audio Anti-Spoofing Dataset) [26], generated using 91 TTS models and covering 38 languages. It expands upon the M-AILABS speech datasets [47], which provide recordings of pristine human speech in eight languages sourced from audiobooks and interviews. The data creation process of MLAAD differs significantly from other existing works that directly utilize human voices as non-generated speech. Specifically, MLAAD labels data as 'benign' and 'spoofed', and for languages not present in M-AILABS, the 'benign' data was created through neural machine translation of existing human speech. MLAAD has demonstrated its effectiveness by enabling the training of deepfake detection models with superior performance compared to other datasets, such as In-The-Wild [25] and FakeOrReal [33] datasets, and its large linguistic diversity enables more robust cross-lingual evaluation and model generalization.

Efforts have also been directed towards audio deepfake detection in low-resource languages. For instance, research has explored languages within the ASEAN region [23]. This preliminary study investigated the development of speech spoofing countermeasures for

[2] https://www.asvspoof.org/
[3] http://addchallenge.cn/add2023

ASEAN languages, specifically detailing datasets for Thai (ThaiSpoof [12]), Indonesian (InaSpoof [2, 22]), Vietnamese (VSASV [14]), and Myanmar (UCSYSpoof [27]). This work highlights several key challenges inherent in the development of robust detection models for low-resource languages, including the limited availability of pristine data sources, the variation in data quality across different languages, and the continuous evolution of high-quality spoofing techniques.

In summary, although existing datasets have contributed significantly to the field of generated voice detection, many remain limited in linguistic diversity and lack sufficient balance across languages.

2.2 Generated Voice Detection Methods

Methods for detecting generated speech can be broadly categorized into two types of approaches: (1) a pipeline architecture consisting of a front-end feature extractor and a back-end classifier, and (2) an end-to-end (E2E) architecture that operates directly on raw audio waveforms while jointly optimizing the feature extraction and classification processes [19, 45]. Benchmark models have primarily been established within the context of research challenges such as the ASVspoof and ADD series, where new systems are evaluated against established baselines.

Early development stages often utilizes the pipeline approach and the front-end feature extractors are typically categorized into hand-crafted features, which served as foundational baselines in earlier challenges, and deep features, which are learned using deep neural networks. Well-known spectral features include linear frequency cepstral coefficients (LFCC) and constant-Q cepstral coefficients (CQCC), both of which have demonstrated strong performance in ASVspoof and ADD challenges [45]. However, most hand-crafted features suffer from design biases due to the limitations of their representations [46]. To overcome this, deep features have been introduced and increasingly adopted. Features extracted from models pre-trained on large-scale speech corpora have shown considerable success and are featured in many top-performing systems in the benchmark challenges [19, 45]. One of the most prominent pre-trained embedding features is derived from XLS-R [4], a variant of wav2vec 2.0, which has demonstrated high effectiveness and robustness in detecting fake audio.

Similarly, the back-end classifier was initially implemented using traditional machine learning methods, most commonly support vector machines, Gaussian mixture models, and random forests. In particular, GMM-based classifiers are frequently used alongside LFCC or CQCC features and serve as baselines in benchmark challenges [40, 45]. More recently, deep learning-based classifiers have significantly outperformed traditional methods due to their superior modeling capabilities [13]. Notable architectures include light convolutional neural network (LCNN), residual network, and graph neural networks (GNN) [19]. One limitation of pipeline approach is their high dependency on extracted features, as information lost during feature extraction is often irretrievable [19]. As a result, E2E architectures have gained increasing attention.

Two prominent E2E models are RawNet [36] and AASIST [16], both of which serve as baselines in the most recent ASVspoof challenge [8]. RawNet employs a fixed bank of sinc filters and residual blocks with gated recurrent units (GRUs) to convert frame-level representations into utterance-level embeddings. It is widely recognized as one of the most reproducible and well-established models in generated audio detection. AASIST, on the other hand, extends RawNet with spectro-temporal heterogeneous graph attention layers to enhance representation learning. It has been noted for outperforming current state-of-the-art E2E models in several evaluation scenarios [45].

Despite extensive efforts, a recurring challenge is the generalization of detection systems beyond the English datasets and controlled benchmark environments [25, 45]. Studies have shown that while many models perform well on standard benchmarks, their effectiveness drops significantly on real-world or cross-language data, indicating overfitting to specific datasets and a lack of robustness to diverse spoofing techniques and acoustic conditions. To address the generalization issue, recent studies have explored both the adaptation of existing models and the development of novel architectures for multilingual audio deepfake detection. For instance, the adversarial-based domain adaptation paradigm has been employed to train models to discriminate between real and fake audio while minimizing reliance on language-specific features [3]. Furthermore, the effectiveness of multilingual speech pre-trained models (PTMs) for audio deepfake detection has also been evaluated across three varied benchmark settings. The findings suggest that multilingual PTMs, when combined with simple downstream networks, outperformed other PTM representations in audio deepfake detection, supporting the hypothesis that linguistic diversity during pre-training enhances robustness [29].

3 Proposed Method

To address the challenges of robust and generalizable audio deepfake detection across diverse languages and spoofing techniques, we propose a hybrid modeling approach that synergistically combines the strengths of end-to-end raw waveform processing with the learned representations from self-supervised learning (SSL) on multilingual data. Our method leverages two prominent architectures: RawNet2 [36] for its efficacy in directly learning discriminative features from raw audio, and AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) [16] for its attention-based mechanism that has shown promise in capturing spoofing artifacts. Furthermore, to enhance the model's ability to generalize to unseen languages and spoofing methods, we integrate a multilingual SSL model to inject language-agnostic representations. Finally, to further improve the robustness of our detection models against various types of noise, we adopted an augmentation method based on RawBoost [35].

Figure 1 presents an overview of our proposed method, which consists of four primary modules. The first module is based on the AASIST model, directly processing raw speech signals. The second module, SSL-RawNet2-AASIST, utilizes the front-end system of SSL-RawNet2 and the AASIST model as its back-end. The third module is the optional RawBoost augmentation module. Finally, the fourth module is the decision model, which determines whether the input signal is pristine or generated speech.

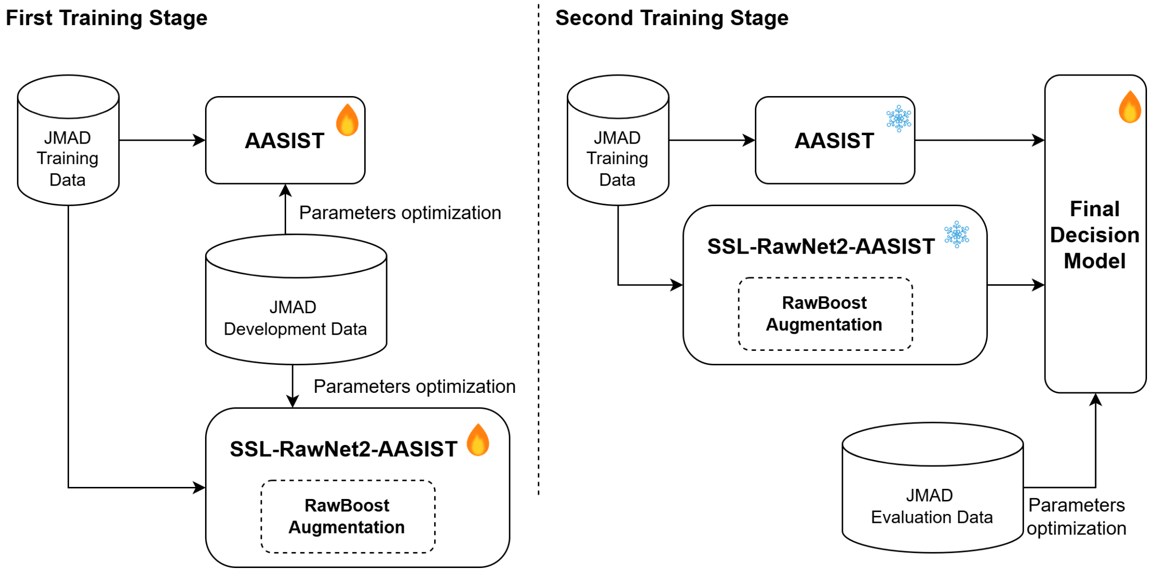

Figure 1: Illustration of our proposed two-stage hybrid method. The process involves two distinct training stages. Models indicated by a fire icon are actively trained in their respective stages, while models marked with a snowflake icon have their weights frozen.

3.1 AASIST module

AASIST is an end-to-end model for audio deepfake detection which builds upon the SincNet layer for efficient front-end filtering and integrates an attention mechanism to focus on the most discriminative temporal segments within the audio [16]. AASIST has shown strong performance in detecting various types of audio spoofs by effectively capturing subtle temporal inconsistencies and artifacts introduced by synthesis and manipulation processes, and it has become a state-of-the-art method, particularly on the ASVspoof 2019 dataset [37]. By including AASIST in our hybrid framework, we aim to leverage its sensitivity to fine-grained spoofing cues.

3.2 SSL-RawNet2-AASIST module

To enhance the cross-lingual generalization capabilities of our hybrid model, we propose to integrate representations learned from a pre-trained multilingual SSL model. SSL models, trained on massive amounts of unlabeled audio data across multiple languages, learn rich contextualized representations that capture linguistic and acoustic commonalities. We integrated the SSL embeddings as feature fusion. Extracting embeddings from an intermediate or final layer of the pre-trained SSL model and concatenating them with the features extracted from RawNet and AASIST before the classification layer. This allows the downstream classifier to leverage both task-specific and language-agnostic representations. We investigated the effectiveness of two multilingual SSL models, such as wav2vec2.0 XLS-R [4] and WavLM [6], based on their performance on multilingual speech tasks and their suitability for feature extraction or fine-tuning.

3.3 RawBoost module

To enhance the robustness of our proposed detection models against various real-world degradations and channel effects, we incorporate an optional data augmentation module based on the RawBoost method [35]. RawBoost is a data-driven augmentation technique specifically designed for raw waveform inputs in spoofing detection. Unlike other augmentation methods that might require external noise recordings or impulse responses, RawBoost operates solely on the existing training data, making it agnostic. The core principle of RawBoost is to simulate variability commonly encountered in real-world scenarios, particularly telephony. It achieves this by applying a combination of several transformations to the raw audio signal, incuding linear and non-linear convolutive noise, impulsive signal-dependent additive noise, and stationary signal-independent additive noise. By randomly applying and parameterizing these transformations during training, RawBoost exposes the models to a wider range of acoustic conditions, forcing them to learn more invariant features that are less susceptible to noise and distortions. This optional module can be selectively applied during the training process to improve the generalization capabilities and real-world performance of our RawNet and AASIST-based detection models, without requiring any external data or significantly increasing the complexity of the training pipeline.

3.4 Decision module

Our proposed hybrid model combines the back-end classifier pathways of its constituent modules through concatenation or late fusion techniques. To obtain the final classification layer for predicting whether the input audio is pristine or generated, we froze the weights of the first and second modules. This final classification layer learns to weight the contributions of the first and second

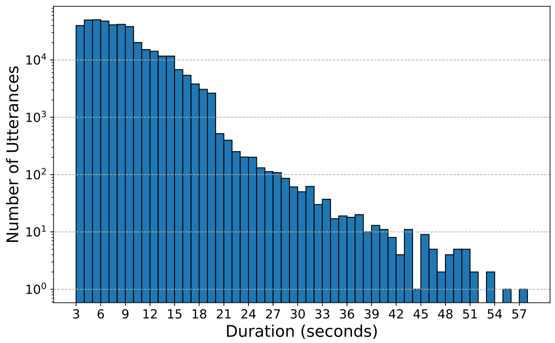

Figure 2: Histogram showing the distribution of utterance durations in the JMAD dataset (All Source) on a logarithmic scale

modules based on the JMAD evaluation data. Subsequently, we perform a grid search to determine the optimal hard final threshold for the decision boundary between pristine and generated signals.

By combining the strengths of end-to-end raw waveform processing, attentive spoofing artifact analysis, language-agnostic representations from multilingual SSL, and optional data augmentation with RawBoost, we aim to develop a robust and highly generalizable audio deepfake detection system capable of effectively addressing the evolving threats in diverse environments.

4 Experiment

The **S**ynthetic **A**udio **F**orensics **E**valuation (SAFE) Challenge is one of initiatives aimed at fostering advancements in the detection of synthetic and manipulated audio. Recognizing the growing sophistication of audio generation and editing techniques, the challenge presents participants with three distinct detection tasks: (***Task 1***) identifying generated audio, (***Task 2***) identifying processed audio, and (***Task 3***) identifying laundered audio. The SAFE challenge is entirely blind (no datasets for training or evaluation will be publicly released). Participants only have access to a limited sample dataset for pilot testing. The submission of detection models was conducted via the Hugging Face[4] platform, and these submissions are evaluated based on balanced accuracy.

4.1 Dataset

Given that the SAFE challenge is entirely blind and does not provide any training or development data, we utilized JMAD (**J**AIST **M**ultilingual **A**udio **D**eepfake) dataset [24]. This dataset comprises a total of 412,021 speech samples spanning 17 languages, representing several of the world's most widely spoken languages. The majority of the dataset was constructed from open-source resources [8, 15, 17, 26, 34, 43, 47], with supplementary contributions from private-source datasets and small portions of internally recorded speech where open resources lacked sufficient quantity or quality of human speech. Parts of the private-source data were drawn from our previous work on spoofing detection in Asian languages [1, 2, 12, 22, 23, 27].

Based on the source of the speech data, the dataset is organized into two main configurations. The *Open Source* subset consists entirely of publicly available corpora and serves as the primary focus to promote transparency and reproducibility. This subset encompasses 15 languages: English (eng), Mandarin (zho), Hindi (hin), Italian (ita), Arabic (arb), Spanish (spa), Polish (pol), German (deu), French (fra), Russian (rus), Portugese (por), Japanese (jpn), Ukranian (ukr), Vietnamese (vie), Thai (tha). In contrast, the *All Source* configuration includes both open-source and private-source data, providing supplementary data for four languages (English, Japanese, Vietnamese, and Thai) and full additional data for two low-resource languages: Indonesian (ind) and Myanmar (mya).

To reflect real-world conditions where machine-generated speech often exceeds human speech in volume, the number of generated utterances in the dataset was set to approximately three times that of pristine utterances, with a larger portion allocated to English and Mandarin in accordance with their global prevalence. All audio samples were constrained to a maximum duration with most samples averaging around 5 seconds, as shown in Fig. 2. To ensure consistency, all audio was standardized to a 16 kHz, mono-channel WAV format and rigorously checked for integrity. The finalized data was then organized into a unified structure with corresponding metadata, and the partitioning is detailed in Subsection 4.4.

4.2 Evaluation Metrics

The SAFE Challenge evaluates submissions based on three main metrics: Pristine Accuracy, Generated Accuracy, and Balanced Accuracy. The '**Pristine**' label denotes authentic audio/ human voice, while '**Generated**' indicates synthetic audio. The definitions are as follows:

$$\text{Pristine Accuracy} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Generated Accuracy} = \frac{TN}{TN + FN} \tag{2}$$

$$\text{Balanced Accuracy} = \frac{\text{Pristine Accuracy} + \text{Generated Accuracy}}{2} \tag{3}$$

In evaluation, true positive (TP) is a correctly identified pristine sample, false positive (FP) is a pristine sample incorrectly labeled as generated, true negative (TN) is a correctly identified generated sample, and false negative (FN) is a generated sample incorrectly labeled as pristine.

In benchmark challenges, Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) often serve as important metrics for audio deepfake detection [8]. EER is the point where the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) of a system are equal. A lower EER generally indicates a more balanced and accurate system, as it signifies a threshold where the trade-off between incorrectly accepting a generated sample as pristine and incorrectly rejecting a pristine sample as generated is minimized. On the other hand, the minimum Detection Cost Function (minDCF) is a more application-aware metric. It considers the costs associated with both false positives and false negatives, as well as the prior

[4] https://huggingface.co/safe-challenge

Table 1: Initial analysis of the RawSpeech-AASIST model, pre-trained on ASVspoof datasets, when evaluated on Task 1 of the SAFE Challenge. The EER-eval (%) column presents the EER achieved on the corresponding ASVspoof evaluation set. Arrows indicates the direction of better performance for each metric ((↑) = higher is better, (↓) = lower is better).

Training data	EER-eval (%) (↓)	Accuracy (%) on SAFE eval. data (task 1)		
		Generated (↑)	Pristine (↑)	Balanced (↑)
ASVspoof2019	0.83	81.00	4.40	42.70
ASVspoof2024	15.20	78.79	16.90	47.84

probability of the target class. By minimizing this cost function over different operating points (thresholds), minDCF provides a measure of the best possible performance a system can achieve under specific operational conditions and cost assumptions.

During model development, we also utilized the Area Under the Receiver Operating Characteristic Curve (AUC) to determine an optimal decision threshold for detection. The AUC provides a measure of the model's ability to distinguish between the pristine and generated classes across various threshold settings, allowing us to select a threshold that balances precision and recall.

Since balanced accuracy determines the leaderboard rankings, making its improvement is our primary objective. EER, minDCF, and AUC serve as secondary objective metrics in our experiments.

4.3 Initial Analysis

This section details our initial experimental setup and findings. As the languages and synthesis techniques in the SAFE challenge are totally unknown, our initial analysis involved submitting the pre-trained RawSpeech-AASIST model [16] on ASVspoof 2019 [39] and trained RawSpeech-AASIST model on ASVspoof 2024 [8]. This initial analysis served to assess the reference baseline performance of a model trained on English data from ASVspoof challenges.

Table 1 presents the performance of the RawSpeech-AASIST model pre-trained on ASVspoof challenge datasets, evaluated on Task 1 of the SAFE Challenge. The 'EER-eval' column indicates the performance of each pre-trained model on the evaluation set of its corresponding data source. The RawSpeech-AASIST model achieved a low EER on ASVspoof 2019, suggesting that this dataset is relatively easy to fit and its evaluation set is not significantly different from the training and development sets. In contrast, the latest ASVspoof 2024, which features a more challenging evaluation set, resulted in poor performance for the pre-trained RawSpeech-AASIST model on its corresponding evaluation set (EER of 15.2%). When these equivalently pre-trained models were evaluated on SAFE Challenge Task 1, the performance was considerably worse. The calculated class-wise accuracy revealed a significant imbalance, with models predominantly classifying input as generated audio. Specifically, the pristine accuracy was extremely low: 4.40% and 16.90% using pre-trained models which trained on ASVspoof 2019 and ASVspoof 2024, respectively. This tendency to over-predict the 'generated' class implies that the SAFE Challenge evaluation data is much more challenging, likely containing multilingual speech and diverse, unseen synthesis techniques to which the pre-trained models have not generalized well.

Table 2: Distribution of the JMAD dataset into training (Train), development (Dev), and evaluation (Eval) sets, as used for model validation during our experiments.

Partition	Subset	#utts.		
		Pristine	Generated	Total
All	Train	44,442	137,723	182,165
	Dev	24,977	90,852	115,829
	Eval	24,547	89,478	114,025
Open	Train	28,981	100,452	129,433
	Dev	18,269	73,050	91,319
	Eval	18,296	73,098	91,394
Clean	Train	42,509	96,522	139,031
	Dev	17,339	59,795	77,134
	Eval	18,296	73,098	91,394

Following this initial analysis, we carefully investigated the training process and the generalization challenges specifically using English data. First, regarding pristine data, ASVspoof challenges predominantly utilized data from single sources, whereas the SAFE Challenge likely encompasses data from various origins. The quality of pristine data in the SAFE Challenge might also vary, extending beyond recordings from controlled settings. Furthermore, the speakers in ASVspoof datasets are likely from similar geographical regions. Based on this investigation, we carefully split JMAD dataset to create a model validation set, the details of which are explained in the subsequent section.

4.4 Model Validation Set

We partitioned the dataset into three subsets—training, development, and evaluation—using a 6:2:2 ratio. The splitting was primarily based on speaker ID and the source dataset. However, we also considered balancing the distribution of attack ID and gender where this information was available. This process ensures that each subset maintains a similar distribution of these attributes. Furthermore, the three subsets are mutually exclusive, with no overlapping audio samples between them.

We also recognized that low-quality audio data could degrade model performance by introducing noise during training. To mitigate this, we created a separate data partition that was pre-filtered using Mean Opinion Score (MOS) values obtained from DNSMOS [32] to purify the training and development sets. This subset, derived from 'All Source' was named as the 'Clean' partition. Specifically, we applied thresholds of 2.6 for pristine-all samples and 2.0 for generated-opensource samples. The lower threshold for generated audio reflects our intention to retain some low-quality synthetic audio, as such artifacts are often present in generated audio due to the synthesis methods. For the evaluation set, we deliberately avoided any filtering to preserve a more diverse set of audio samples, ensuring that the final evaluation reflects robustness.

Table 2 shows the distribution of the training, development, and evaluation sets for all partitioned data used in our experiments. During pre-analysis, we identified a very small number (less than 5) of problematic utterances which have been temporarily removed for the current experiments. In the near future, we plan to investigate the reasons for these issues and implement a fix.

4.5 Model Configurations

This section shows some configurations of our primary modules and the key hyperparameters we tuned during experimentation.

4.5.1 Model Parameterization. AASIST Module: This model directly processes raw audio waveforms. Its architecture follows the standard AASIST configuration[5] [16], which includes a SincNet layer for front-end filtering, followed by convolutional blocks, self-attention layers, and a final classification layer. The primary architectural parameters are the number and size of convolutional filters, the number of attention heads, and the dimensions of the hidden layers. These were largely kept consistent with the original AASIST paper for a strong baseline.

SSL-RawNet2-AASIST Module: This module integrates self-supervised learning (SSL) embeddings. We experimented with the multilingual pre-trained models XLS-R (300M) [4], XLS-R (1B) [4], and WavLM [6]. The audio was first processed by the SSL model to extract frame-level embeddings. These embeddings were then used in two ways: either directly fed into a modified AASIST backend (referred to as SSL-AASIST (Direct)), or concatenated with learned features from a RawNet2 front-end before being processed by the AASIST architecture (SSL-RawNet2-AASIST). We also experimented with data augmentation using the RawBoost algorithm, with our base implementation following the publicly available code[6] [35]. The key parameters explored within this module include the choice of the SSL model and the strategy for integrating its embeddings, such as the layer from which embeddings are extracted and the dimensions of any intermediate projection layers.

4.5.2 Hyperparameter Tuning. During our experiments, we explored the impact of several key training hyperparameters and architectural choices, including padding, learning rate, and objective functions.

The AASIST model requires a fixed input dimension, so all audio clips were either truncated or padded before being fed into the network. To handle variable-length input audio, we experimented with different padding strategies and input lengths. Initially, we used a fixed input length of 64,600 samples, corresponding to approximately four seconds of audio at a 16 kHz sampling rate. Shorter audio clips were padded with repetition, while longer ones were truncated. We also experimented with input lengths of 160,000 samples (around 10 seconds), 192,000 samples (around 12 seconds), and 240,000 samples (around 15 seconds). The fixed input length was a key hyperparameter we tuned based on the development set performance. The extension of the padding enables model to access longer temporal features. This process significantly improved the model's classification performance on longer audio clips.

We primarily employed the standard Categorical Cross-Entropy (CCE) loss for binary classification (pristine vs. generated). The mathematical function of CCE loss ($\mathcal{L}_{CCE}$) is expressed as follows.

$$\mathcal{L}_{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}) \tag{4}$$

where N is the number of samples, C is the number of classes, y_{ij} is the indicator (0 or 1) if sample i belongs to class j, and p_{ij} is the predicted probability that sample i belongs to class j.

Additionally, we investigated the use of Weighted Cross-Entropy (WCE) to address the class imbalance issue between pristine and generated samples. The weights assigned to each class in the WCE loss ($\mathcal{L}_{WCE}$) were tuned based on the class distributions in the training data.

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j y_{ij}\log(p_{ij}) \tag{5}$$

where w_j is the weight assigned to class j.

During training, we optimized our models using the AdamW optimizer, and we experimented with learning rates of 10^{-4}, 10^{-5}, and 10^{-6}. Our primary objective function for training was the CCE loss (Eq. 4) or WCE loss (Eq. 5), to minimize the classification error between pristine and generated audio. The final model selection and evaluation of generalization capabilities were primarily guided by the EER achieved on our held-out validation data, as a lower EER signifies better overall detection performance by balancing false positives and false negatives.

4.6 Results and Analysis

The primary objective of our experiments was to evaluate the generalization capabilities of our proposed models for audio deepfake detection, particularly across diverse languages and unseen spoofing techniques, as relevant to the SAFE Challenge. To achieve this, we employed a two-stage evaluation process, first assessing performance on JMAD dataset and subsequently on the blind SAFE Challenge evaluation set.

4.6.1 Overall Performance on JMAD Dataset. Before conducting a comprehensive evaluation, we trained our models on three different subsets of the dataset (JMAD-All, JMAD-Open, and JMAD-Clean) as illustrated in Fig. 3. We then selected five representative modules in our proposed method: RawSpeech-AASIST, SSL-AASIST, SSL-RawNet2-AASIST, SSL-RawNet2-AASIST with RawBoost augmentation. This comparison highlights the improvements in detection performance achieved by integrating multilingual self-supervised learning representations and by combining different architectural strengths. Within this context, we also investigated the impact of different SSL models, padding strategies, and other parameters, as detailed in Subsection 4.5. Table 3 presents the results of the top five representative systems, showcasing various parameter settings, assessed on our JMAD-Open dataset.

As overall results, XLSR-RawNet2-AASIST with 10 seconds padding gave the best performance in the single module evaluation. During our experiments with multilingual SSL models, we considered XLS-R with both 300M and 1B parameters. However, we observed that the XLS-R 1B model presented challenges in achieving high objective scores during the initial training epochs and demanded significantly more computational time and resources. Consequently, for practical reasons and to maintain a feasible experimental scope, we primarily focused our in-depth analyses on the XLS-R 300M model. We also explored the WavLM model. While it outperformed the RawSpeech-AASIST model on the JMAD-Open evaluation set, its performance on the SAFE evaluation set was

[5] https://github.com/clovaai/aasist

[6] https://github.com/TakHemlata/SSL_Anti-spoofing

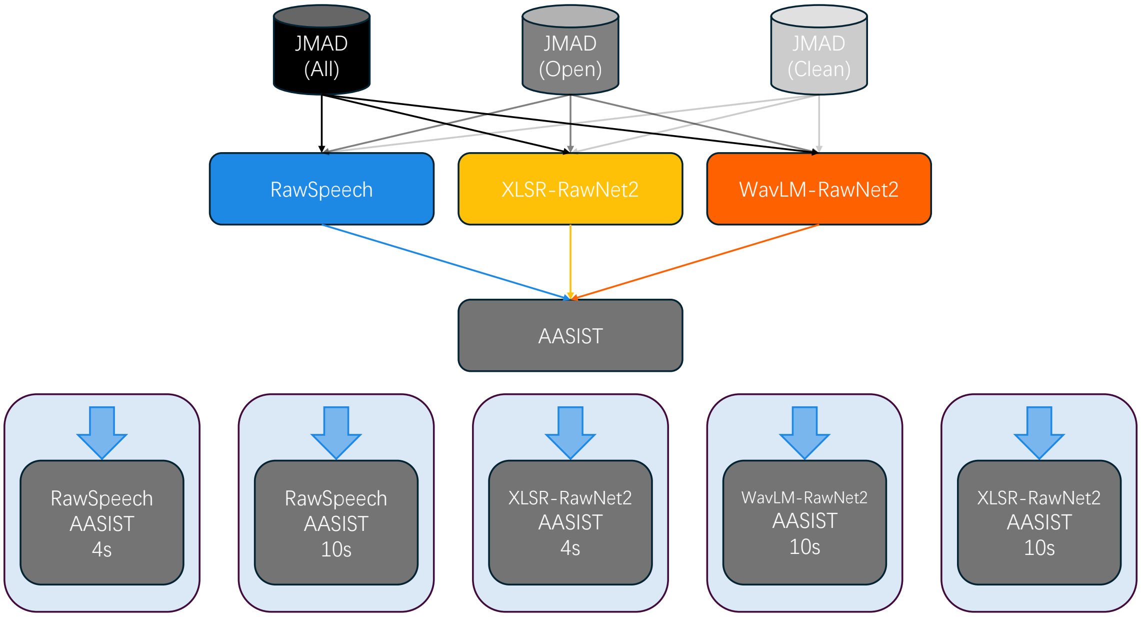

Figure 3: Models trained using various JMAD subsets.

Table 3: Overall performance evaluation on the JMAD-Open dataset. The table presents the top five representative models with different parameter and architectural configurations.

Front-end	Back-end	Padding	Augmentation	Label	Accuracy (%)	Balanced Acc. (%)	AUC (%)	EER (%)	minDCF
RawSpeech	AASIST	4s	No	Generated	98.97	79.14	93.51	12.39	0.22
				Pristine	59.32				
RawSpeech	AASIST	10s	No	Generated	98.84	78.01	92.73	13.75	0.24
				Pristine	57.17				
XLSR-RawNet2	AASIST	4s	RawBoost	Generated	99.79	90.65	97.86	4.67	0.07
				Pristine	81.52				
WavLM-RawNet2	AASIST	10s	RawBoost	Generated	98.06	80.09	95.22	10.99	0.26
				Pristine	62.12				
XLSR-RawNet2	AASIST	10s	RawBoost	Generated	99.88	91.14	98.01	4.21	0.06
				Pristine	82.39				

poor (around 50% balanced accuracy). Therefore, we also decided to exclude WavLM from further in-depth evaluation.

Regarding the impact of padding length (4s vs. 10s) in the RawSpeech-AASIST and SSL-RawNet2-AASIST models, our evaluation on JMAD dataset yielded almost similar results across most metrics for both durations. However, the evaluation on the SAFE Challenge Task 1 data revealed a significant improvement (an increase of nearly 10% in balanced accuracy) when using 10s padding compared to 4s. This trend is also consistent with the behavior observed in the SSL-RawNet2-AASIST model. While the difference was not substantial on our evaluation set, this clearly indicates a different tendency on the unseen SAFE evaluation data. Particularly, utilizing longer audio segments on RawSpeech-AASIST model achieving up to 66% average accuracy on the challenge test set (task 1). On the other hand, the XLSR-RawNet2-AASIST model reached a balanced accuracy of 60.72% in task 1. Therefore, for the subsequent model combination, we focus on using the longer padding (≥ 10 seconds).

4.6.2 Analysis of Detection Accuracy by Language and Source Dataset.

4.6.2 Analysis of Detection Accuracy by Language and Source Dataset. To gain deeper insights into the models' performance across languages, we conducted a detailed analysis of the detection accuracy using our best individual model based on overall performance on JMAD-Open: XLSR-RawNet2-AASIST. This analysis across the 15 languages in our JMAD-Open dataset reveals potential language-specific biases and varying difficulty in detecting generated audio from existing speech synthesis technologies.

Figure 4 displays the distribution of countermeasure scores for pristine and generated classes across each language. As illustrated, most languages exhibit clear separation between the two classes, achieving a near 100% F1 score, with the notable exception of Mandarin Chinese (zho). For the pristine class, zho showed the lowest F1 score at approximately 76%. Conversely, for the generated class,

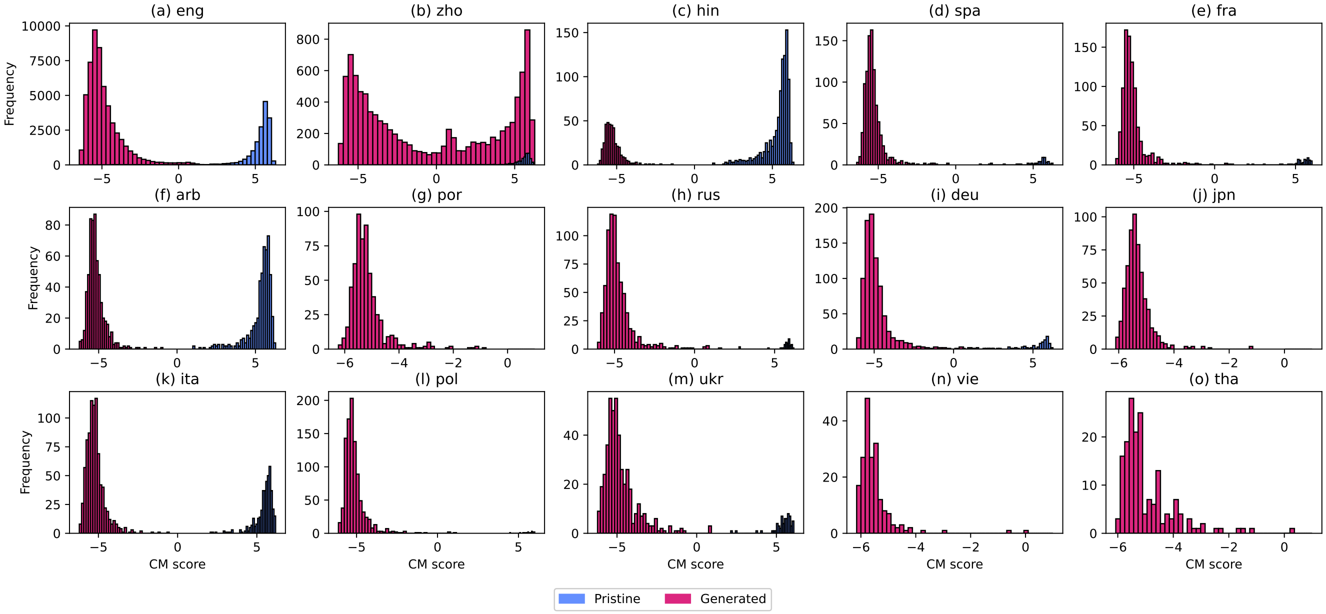

Figure 4: Countermeasure (CM) score distributions from the XLSR-RawNet2-AASIST method on the JMAD-Open evaluation dataset, categorized by language. These plots highlight the discrimination achieved by the method. Some individual datasets within JMAD contain only one type of data (pristine or generated).

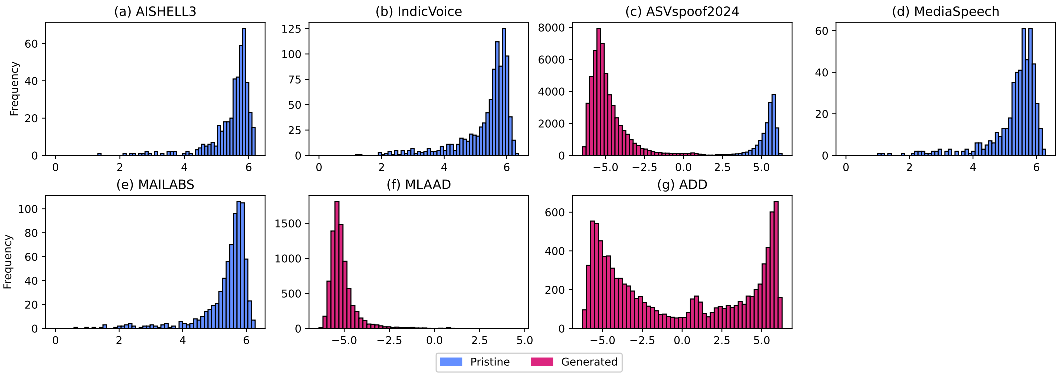

Figure 5: CM score distributions from the XLSR-RawNet2-AASIST method on the JMAD-Open evaluation dataset, categorized by dataset source.

we observed particularly poor detection, with an F1 score of only around 18.14%. We hypothesize that this could be due to the data for the Chinese portion originating from the ADD 2022 corpus [43], which is known to contain a significant amount of noisy and low-quality data. This characteristic likely also explains why many systems struggled to achieve high detection rates in the ADD challenges, with the lowest EER achieved by top performers in Track 1 being just 21.7%.

Furthermore, we examined the model's performance based on the source datasets within our training corpus, aiming to understand its ability to generalize beyond specific data origins and synthesis methods. Figure 5 displays the distribution of log-likelihood scores for pristine and generated classes for each source dataset. The results show a similar trend to the language analysis, likely because some languages are predominantly sourced from a particular dataset. Most datasets achieved high accuracy (nearly 100%) in distinguishing both pristine and generated classes, with the exception of ADD, which hovered around 76% in F1 score of generated class. It is important to note that we treated AISHELL-3 (only pristine data) and ADD (only generated data) separately, which differs from the actual ADD challenge that includes AISHELL pristine data. This analysis helps pinpoint areas where the model performs well and where further improvements may be necessary.

4.6.3 Robustness Against Unseen Generated, Processed, and Laundered Data. Finally, we evaluated the robustness of our models

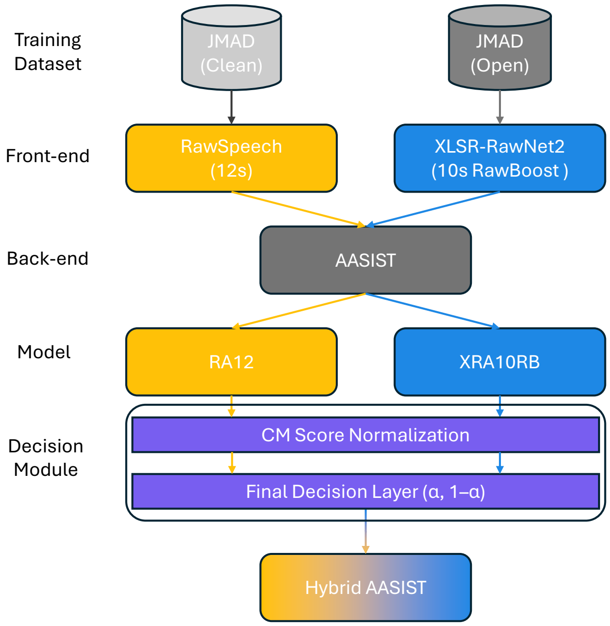

Figure 6: Hybrid prediction model creation: Two top-performing individual models are selected for combination.

on SAFE evaluation which includes unseen generated audio (Task 1), processed audio (Task 2) and laundered audio (Task 3). This involved assessing the detection accuracy on audio samples that have undergone various transformations, such as the application of noise or filtering, to mimic real-world degradation. The results in this section will demonstrate how robust our proposed hybrid method on unseen evaluation data.

Table 4 presents the detection accuracy of the best performing RawSpeech-AASIST module and XLSR-RawNet2-AASIST module. Generally, the non-hybrid AASIST model trained solely on RawSpeech yielded higher performance than those incorporating SSL models without augmentation (Task 1). However, when the audio underwent processing such as codec changes or noise addition, or was "laundered" through artifact reduction methods, the XLSR-RawNet2-AASIST model demonstrated better performance (Task 2 and Task 3). Subsequently, adjusting the classification threshold resulted in an average accuracy improvement of approximately 2%.

We implemented our proposed hybrid framework, illustrated in Fig. 6, by combining the top-performing RawSpeech-AASIST (RA12, trained on JMAD-Open) and XLSR-RawNet2-AASIST (XRA1ORB, trained on JMAD-Clean) modules. Following the procedure detailed in Subsection 3.4, we first normalized the countermeasure scores from each module and then fused them using a weighted decision layer. This adaptive fusion mechanism allows the hybrid system to effectively leverage the complementary strengths of both individual models to generate a final prediction.

This hybrid approach resulted in further gains in balanced accuracy, although this improvement came with the expected trade-off of increased inference time. Specifically, incorporating the XLSR model in the front-end processing led to a substantial increase in inference time, approximately tripling it. Our best-performing hybrid method achieved balanced accuracies of 72.91%, 73.39%, and 65.85% for Tasks 1, 2, and 3, respectively. As we currently lack access to the audio data of the SAFE Challenge evaluation set, a more detailed analysis of the results beyond balanced accuracy is not possible at this stage. We plan to conduct a more comprehensive evaluation once the official challenge outcomes are made available.

5 Conclusion

This paper addressed the critical challenge of robust multilingual audio deepfake detection, particularly within the SAFE Challenge framework. We utilized multilingual and multi-source speech corpus (17 languages) to develop robust detection models. Our proposed hybrid detection model strategically combined the strengths of end-to-end models (i.e., RawNet, AASIST) and the language-agnostic representations learned from multilingual self-supervised learning. We also explored the benefits of data augmentation with RawBoost to enhance robustness against real-world noise.

Our experimental results, including a detailed analysis of generalization across languages and spoofing algorithms, highlighted the limitations of models trained on monolingual datasets when faced with diverse, unseen data, as often encountered in real-world scenarios (represented by the SAFE Challenge evaluation data). The performance of our hybrid model demonstrated promising improvements in overall detection, suggesting the effectiveness of integrating multilingual SSL representations and leveraging a diverse language training corpus.

In conclusion, this work offers valuable insights into multilingual deepfake detection and introduces a promising hybrid modeling strategy to address the increasing global challenge. Future research will build upon these findings, focusing on further analysis and enhancement of the generalization capabilities of advanced detection methods, particularly in the context of multilingual scenarios highlighted by our evaluation.

Data Availability Statement

The **J**AIST **M**ultilingual **A**udio **D**eepfake (JMAD) dataset that utilized for this study includes data from both publicly available and private sources. A list of the publicly available datasets used, along with relevant citations can be found in Subsection 4.1. Due to the inclusion of proprietary data from collaborative projects, the full dataset cannot be made publicly available. However, aggregated statistics and analyses of the dataset are provided within the paper to support our findings. Researchers interested in replicating our work are encouraged to utilize the described publicly available resources.

Acknowledgments

This work was partially supported by JSPS KAKENHI (25K21245, 23K18491, and 25H01139). Additionally, we extend our sincere thanks to our collaborators in the ASEAN IVO project titled "Spoof Detection for Automatic Speaker Verification"(www.nict.go.jp/en/asean_ivo) for their prior collaborative work which contributed to this study.

Table 4: Comparison of the detection performance of our best two individual modules and their hybrid fusion on the SAFE Challenge evaluation (Task 1: generated, Task 2: processed, Task 3: laundered).

Task	Hybrid	ID	Model Description	Partition	Accuracy (%)			Time (s)
					Generated	Pristine	Balanced	
1	No	RA12	RawSpeech-AASIST (12s)	Clean	64.86	67.20	66.03	162.98
	No	XRA10RB	XLSR-RawNet2-AASIST (10s) with augmentation	Open	49.00	72.45	60.72	440.13
	Yes	HYB2	RA12+XRA10RB	Mix	53.57	92.25	**72.91**	585.26
2	No	RA12	RawSpeech-AASIST (12s)	Clean	39.14	72.30	55.72	178.67
	No	XRA10RB	XLSR-RawNet2-AASIST (10s) with augmentation	Open	61.77	80.40	71.08	599.48
	Yes	HYB2	RA12+XRA10RB	Mix	69.62	77.15	**73.39**	822.98
3	No	RA12	RawSpeech-AASIST (12s)	Clean	48.86	68.20	58.53	161.37
	No	XRA10RB	XLSR-RawNet2-AASIST (10s) with augmentation	Open	47.57	80.40	63.99	446.27
	Yes	HYB2	RA12+XRA10RB	Mix	54.86	76.85	**65.85**	570.47

References

[1] Aulia Adila, Candy Olivia Mawalim, and Masashi Unoki. 2024. Detecting Spoof Voices in Asian Non-Native Speech: An Indonesian and Thai Case Study. In *Proc. of APSIPA ASC 2024, December 3-6, 2024*. IEEE, Macau, 1–6. doi:10.1109/APSIPAASC63619.2025.10848707

[2] Sarah Azka Arief, Candy Olivia Mawalim, and Dessi Puji Lestari. 2024. Indonesian Speech Anti-Spoofing System: Data Creation and Convolutional Neural Network Models. In *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE, Singapore, 1–6. doi:10.1109/ICAICTA63815.2024.10763091

[3] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring Audio Deepfake Detection Capability across Languages. In *Proc. of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. ACM, New York, NY, USA, 2033–2044. doi:10.1145/3543507.3583222

[4] 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 *Proc. of Interspeech 2022, September 18-22, 2022*. ISCA, Incheon, Korea, 2278–2282. doi:10.21437/INTERSPEECH.2022-143

[5] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Proc. of O-COCOSDA 2017*. IEEE, South Korea, 1–5. doi:10.1109/ICSDA.2017.8384449

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* 16, 6 (2022), 1505–1518. doi:10.1109/JSTSP.2022.3188113

[7] Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. 2022. Open Challenges in Synthetic Speech Detection. In *Proc. of WIFS 2022, December 12-16, 2022*. IEEE, Shanghai, China, 1–6. doi:10.1109/WIFS55849.2022.9975433

[8] Héctor Delgado, Nicholas Evans, Jee-weon Jung, Tomi Kinnunen, Ivan Kukanov, Kong Aik Lee, Xuechen Liu, Hye-jin Shim, Md Sahidullah, Hemlata Tak, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. 2024. *ASVspoof 5 Evaluation Plan*. Technical Report. ASVspoof consortium. http://www.asvspoof.org/

[9] Héctor Delgado, Massimiliano Todisco, Md. Sahidullah, Nicholas W. D. Evans, Tomi Kinnunen, Kong-Aik Lee, and Junichi Yamagishi. 2018. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In *Proc. of Odyssey 2018: The Speaker and Language Recognition Workshop*, Anthony Larcher and Jean-François Bonastre (Eds.). ISCA, Les Sables d'Olonne, France, 296–303. doi:10.21437/ODYSSEY.2018-42

[10] Helen Fraser, Vincent Aubanel, Robert C. Maher, Candy Olivia Mawalim, Xin Wang, Peter Počta, Emma Keith, Gérard Chollet, and Karla Pizzi. 2024. Forensic Speech Enhancement: Toward Reliable Handling of Poor-Quality Speech Recordings Used as Evidence in Criminal Trials. *journal of the audio engineering society* 72 (may 2024), 748–753. Issue 11.

[11] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 2021. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario. In *Proc. of Interspeech 2021*. ISCA, Czechia, 3665–3669. doi:10.21437/Interspeech.2021-1397

[12] Kasorn Galajit, Thunpisit Kosolsriwiwat, Masashi Unoki, Candy Olivia Mawalim, Pakinee Aimmanee, Waree Kongprawechnon, Win Pa Pa, Anuwat Chaiwongyen, Teeradaj Racharak, Surasak Boonkla, Hayati Yassin, and Jessada Karnjana. 2023. ThaiSpoof: A Database for Spoof Detection in Thai Language. In *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, Thai, 1–6. doi:10.1109/iSAI-NLP60301.2023.10354956

[13] Alan Godoy, Flávio Olmos Simões, José Augusto Stuchi, Marcus de Assis Angeloni, Mário Uliani Neto, and Ricardo Paranhos Velloso Violato. 2015. Using Deep Learning for Detecting Spoofing Attacks on Speech Signals. https://api.semanticscholar.org/CorpusID:13962874

[14] V. Hoang, V.T. Pham, H.N. Xuan, P. Nhi, P. Dat, and T.T.T. Nguyen. 2024. VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification. In *Proc. Interspeech 2024*. ISCA, Kos Island, Greece, 5 pages. doi:10.21437/Interspeech.2024-1972

[15] Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, Pratyush Kumar, and Mitesh Khapra. 2024. IndicVoices: Towards building an Inclusive Multilingual Speech Dataset for Indian Languages. In *Findings of the Association for Computational Linguistics: ACL 2024*. ACL, Bangkok, Thailand, 10740–10782. doi:10.18653/v1/2024.findings-acl.639

[16] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas W. D. Evans. 2022. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In *Proc. of ICASSP 2022, 23-27 May 2022*. IEEE, Virtual and Singapore, 6367–6371. doi:10.1109/ICASSP43922.2022.9747766

[17] Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. MediaSpeech: Multilanguage ASR Benchmark and Dataset. arXiv:2103.16193 https://arxiv.org/abs/2103.16193

[18] Bo Li, Yu Zhang, Tara N. Sainath, Yonghui Wu, and William Chan. 2019. Bytes Are All You Need: End-to-end Multilingual Speech Recognition and Synthesis with Bytes. In *Proc. of ICASSP 2019, May 12-17, 2019*. IEEE, Brighton, United Kingdom, 5621–5625. doi:10.1109/ICASSP.2019.8682674

[19] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A Survey on Speech Deepfake Detection. *ACM Comput. Surv.* 57, 7, Article 165 (Feb. 2025), 38 pages. doi:10.1145/3714458

[20] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE ACM Trans. Audio Speech Lang. Process.* 31 (2023), 2507–2522. doi:10.1109/TASLP.2023.3285283

[21] Bartlomiej Marek, Piotr Kawa, and Piotr Syga. 2024. Are audio DeepFake detection models polyglots? doi:10.48550/ARXIV.2412.17924 arXiv:2412.17924

[22] Candy Olivia Mawalim, Sarah Azka Arief, and Dessi Puji Lestari. 2025. InaSAS: Benchmarking Indonesian Speech Antispoofing Systems. *APSIPA Transactions on Signal and Information Processing* (2025). Accepted.

[23] Candy Olivia Mawalim, Kasorn Galajit, Dessi Puji Lestari, Win Pa Pa, and Masashi Unoki. 2025. Challenges in Speech Spoofing Countermeasures for Southeast Asian Languages. ASJ Spring Meeting 2025.

[24] Candy Olivia Mawalim, Yutong Wang, Aulia Adila, Shogo Okada, and Masashi Unoki. 2025. Multilingual Audio Deepfakes Dataset for Robust and Generalizable Detection. https://candyolivia.github.io/assets/pdf/paper/JMADv1.pdf Pre-release.

[25] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Proc. of Interspeech 2022, September 18-22, 2022*, Hanseok Ko and John H. L. Hansen (Eds.). ISCA, Incheon, Korea, 2783–2787. doi:10.21437/INTERSPEECH.2022-108

[26] Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2025. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. arXiv:2401.09512 [cs.SD] https://arxiv.org/abs/2401.09512

[27] Hay Mar Soe Naing, Win Pa Pa, Aye Mya Hlaing, Myat Aye Aye Aung, Kasorn Galajit, and Candy Olivia Mawalim. 2024. UCSYSpoof: A Myanmar Language

Dataset for Voice Spoofing Detection. In *Proc. of O-COCOSDA 2024, October 17-19, 2024*. IEEE, Hsinchu City, Taiwan, 1–5. doi:10.1109/O-COCOSDA64382.2024.10800220

[28] Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly, and Simon King. 2024. Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023. *Comput. Speech Lang.* 90 (2024), 101747. doi:10.1016/j.csl.2024.101747

[29] Orchid Chetia Phukan, Gautam Siddharth Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024. Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake. In *Findings of the ACL: NAACL 2024*. ACL, Mexico City, Mexico, 2496–2506. doi:10.18653/V1/2024.FINDINGS-NAACL.160

[30] 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 (2024), 97:1–97:52. https://jmlr.org/papers/v25/23-1318.html

[31] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. of Interspeech 2020, October 25-29, 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng (Eds.). ISCA, Virtual Event, Shanghai, China, 2757–2761. doi:10.21437/INTERSPEECH.2020-2826

[32] Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2022. DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *Proc. of ICASSP 2022, 23-27 May 2022*. IEEE, Virtual and Singapore, 886–890. doi:10.1109/ICASSP43922.2022.9746108

[33] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019*. IEEE, Timisoara, Romania, 1–10. doi:10.1109/SPED.2019.8906599

[34] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Proc. of Interspeech 2021*. ISCA, Czechia, 2756–2760. doi:10.21437/Interspeech.2021-755

[35] Hemlata Tak, Madhu R. Kamble, Jose Patino, Massimiliano Todisco, and Nicholas W. D. Evans. 2022. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *Proc. of ICASSP 2022, 23-27 May 2022*. IEEE, Virtual and Singapore, 6382–6386. doi:10.1109/ICASSP43922.2022.9746213

[36] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas W. D. Evans, and Anthony Larcher. 2021. End-to-End anti-spoofing with RawNet2. In *Proc. of ICASSP 2021*. IEEE, Toronto, ON, Canada, 6369–6373. doi:10.1109/ICASSP39728.2021.9414234

[37] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Tomi H. Kinnunen, and Kong Aik Lee. 2019. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Proc. of Interspeech 2019, September 15-19, 2019*. ISCA, Graz, Austria, 1008–1012. doi:10.21437/INTERSPEECH.2019-2249

[38] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md. Sahidullah, Tomi Kinnunen, Nicholas W. D. Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. doi:10.48550/ARXIV.2408.08739

[39] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas W. D. Evans, Md. Sahidullah, Ville Vestman, Tomi H. Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, and Zhenhua Ling. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* 64 (2020), 101114. doi:10.1016/J.CSL.2020.101114

[40] Zhizheng Wu, Nicholas W. D. Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.* 66 (2015), 130–153. doi:10.1016/J.SPECOM.2014.10.005

[41] Zhizheng Wu, Tomi Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. of INTERSPEECH 2015, September 6-10, 2015*. ISCA, Dresden, Germany, 2037–2041. doi:10.21437/INTERSPEECH.2015-462

[42] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md. Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas W. D. Evans, and Héctor Delgado. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. https://arxiv.org/abs/2109.00537

[43] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. 2022. ADD 2022: the first Audio Deep Synthesis Detection Challenge. In *Proc. of ICASSP 2022, 23-27 May 2022*. IEEE, Virtual and Singapore, 9216–9220. doi:10.1109/ICASSP43922.2022.9746939

[44] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. 2023. ADD 2023: the Second Audio Deepfake Detection Challenge. In *Proc. of the Workshop on Deepfake Audio Detection and Analysis co-located with 32th IJCAI 2023*, Vol. 3597. CEUR-WS.org, Macao, China, 125–130. https://ceur-ws.org/Vol-3597/paper21.pdf

[45] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio Deepfake Detection: A Survey. doi:10.48550/ARXIV.2308.14970 arXiv:2308.14970

[46] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. {LEAF}: A Learnable Frontend for Audio Classification. In *Proc. of ICLR 2021, Virtual Event*. ICLR, Austria, 16 pages. https://openreview.net/forum?id=jM76BCb6F9m

[47] İmdat Celeste. 2020. The M-AILABS Speech Dataset. https://github.com/imdatceleste/m-ailabs-dataset. Accessed: April 2025.