IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 February 2026, accepted 11 March 2026, date of publication 3 April 2026, date of current version 16 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3680508

RESEARCH ARTICLE

Multilingual Deepfake Speech Dataset for Robust and Generalizable Detection

CANDY OLIVIA MAWALIM, (Member, IEEE), YUTONG WANG, AULIA ADILA, SHOGO OKADA, (Member, IEEE), AND MASASHI UNOKI, (Member, IEEE)
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan

Corresponding author: Candy Olivia Mawalim (candylim@jaist.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 25K21245, Grant 23K18491, and Grant 25H01139; and in part by the Japan Science and Technology Agency (JST) Program for Co-Creating Startup Ecosystem under Grant JPMJSF2318.

ABSTRACT The rise of sophisticated technologies capable of generating realistic synthetic human speech has introduced significant security challenges in voice-based applications. These advances in speech generation have enabled malicious actors to produce convincing speech deepfakes, undermining the reliability of speaker verification systems and digital communication. Despite growing interest in deepfake speech detection, many existing datasets are limited in linguistic diversity and fail to capture the complexity of real-world scenarios, thus constraining model generalization. In this work, we introduce the JAIST Multilingual Deepfake Speech (JMDS) dataset, a large-scale, multilingual, and multi-source corpus designed to support the development and evaluation of robust and generalizable deepfake speech detection systems. Covering 17 languages and comprising over 350 000 utterances, JMDS incorporates a wide range of deepfake generation methods and includes both human (pristine) and machine-generated (spoofed) speech, sourced from publicly available corpora and a curated subset of private data. We provide detailed analyses of utterance duration, generation techniques, and audio quality, along with comprehensive evaluations across multiple model architectures and configurations. Cross-dataset evaluations are also conducted to assess the generalization capabilities of detection models across diverse languages and data domains. This study contributes to a deeper understanding of the limitations and opportunities in current detection systems, ultimately paving the way for more resilient and linguistically inclusive countermeasures.

INDEX TERMS Deepfake speech detection, generalizability, multilingual dataset.

I. INTRODUCTION

The advancement of deep learning technologies has significantly improved the quality of generated content across various modalities. Among these, speech synthesis technologies such as text-to-speech (TTS) and voice conversion (VC) have enabled a wide range of beneficial applications, particularly in human communication. However, they have also introduced serious potential threats to social security and political stability when misused for malicious purposes [1]. One such threat is the use of deepfake speech, which refers to speech that has been digitally altered using artificial intelligence (AI) [1], with the intent to deceive humans. This can lead to harmful consequences such as fraud and the spread of misinformation. Another concern involves biometric identification systems, where manipulated speech is used to bypass automatic speaker verification (ASV) systems [2].

An emerging defense strategy against deepfake speech is spoofing detection, which aims to distinguish between human (pristine) and machine-generated (spoofed) speech. Developing a reliable detection model presents several challenges. First, the continuous evolution of speech synthesis techniques, including diverse model architectures and training algorithms, means that detection models must be able to generalize effectively across a wide variety of attacks. This highlights the importance of training detection systems on realistic and representative datasets to ensure robustness.

Furthermore, as speech generation technologies expand into multilingual settings [3], [4], [5], it becomes increasingly critical to advance spoofing detection capabilities beyond high-resource languages to ensure inclusive protection. While

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

audio data formats are typically common (e.g., PCM, MP3), enabling theoretical cross-lingual detection (training on Language A, testing on Language B), performance degrades significantly due to feature space distribution shifts [6].

The degradation arises because the subtle, system-specific artifacts introduced by a deepfake generator (the ''spoofing signature'') are often learned by a detector in the context of the training language's specific phonetic and prosodic characteristics. When the model encounters a new language, the altered feature space (e.g., new phonemes, different pitch contours, variations in vocal tract filtering) can obscure or shift the learned deepfake signature, leading to high false alarm rates or missed detections. This drop in cross-lingual accuracy is particularly pronounced for low-resource languages, where the scarcity of properly labeled genuine and spoofed data exacerbates the challenge. Most existing benchmarks are English-centric, forcing models to generalize from an over-represented language space to entirely novel acoustic and linguistic domains, resulting in unreliable real-world performance. Addressing this scarcity of diverse, high-quality, labeled multilingual deepfake data is thus essential for building robust deepfake detection systems.

In response to these challenges, we introduce the **J**AIST **M**ultilingual **D**eepfake **S**peech (JMDS) dataset, an initiative designed to facilitate the development of robust detection models capable of distinguishing human from machine-generated speech across multilingual and multi-source conditions. The JMDS dataset comprises a comprehensive multilingual speech corpus compiled from multiple open-source resources and small parts of internally collected data, rigorously curated to ensure both quality and representativeness. Spanning 17 languages and encompassing a broad spectrum of synthesis methods, recording conditions, and speaker demographics, the JMDS dataset is explicitly constructed to improve model generalization beyond what existing resources currently support.

To assess the utility of our dataset, we present a comprehensive evaluation encompassing several key aspects. First, we evaluate the inherent quality of the dataset, establishing its foundational fitness as a high-fidelity resource for deepfake speech research. Second, we quantify the detection performance using benchmark methods specifically on the JMDS dataset. This establishes critical performance metrics and an initial benchmark for future research leveraging this resource. Finally, we conduct a cross-dataset evaluation to demonstrate its potential for training robust models capable of generalizing to unseen data.

II. RELATED WORK

The most widely adopted dataset for advancing the development of deepfake speech detection systems is provided by the ASVspoof Challenge series [2], [7], [8], [27], [28], which has served as a benchmark primarily for English. Earlier editions focused on spoofing attacks targeting automatic speaker verification (ASV) systems, while more recent editions have expanded to include standalone countermeasures independent of ASV. The latest edition, ASVspoof 5 [28], is built on the MLS [29] corpus and incorporates adversarial attacks applied to spoofed utterances generated using various TTS, vocoder, and VC algorithms. It also introduces codec simulation to reflect real-world audio transmission scenarios.

Another well-known initiative is the ADD Challenge [12], [15], which addresses more complex real-world detection scenarios. The first edition focused on detecting low-quality and partially fake audio [12], while the second edition expanded to include manipulation localization and generation method identification [15]. The ADD dataset is derived from Mandarin corpora, AISHELL-1 [30], AISHELL-3 [31], and AISHELL-4 [32], and contains samples generated using a range of TTS and VC models, though the specific models are not publicly disclosed.

Filling the gap, CFAD [13] was introduced as the public Mandarin standard dataset for fake audio detection under noisy and transcoding conditions. It consists of three dataset versions: clean, noisy, and codec. Human speech samples were sourced from both open datasets and self-recorded data (six sources in total). It incorporates 12 types of fake speech, 11 of which are generated using synthesis methods with different vocoders, and one that is partially fake and obtained by clipping and splicing.

Additionally, several English-language corpora have been proposed to advance research in deepfake speech detection. The Fake or Real (FoR) dataset [9], introduced in 2019, includes samples generated using a combination of open-source and commercial tools. The In-the-Wild dataset [14], designed to evaluate model generalization in real-world conditions, comprises found speech recordings of celebrities and politicians, with approximately half of the recordings being deepfakes. Another notable resource is the DFADD dataset [20], which contains deepfake speech generated using five state-of-the-art diffusion and flow-matching TTS models. To address the growing challenges posed by zero-shot TTS systems, the Cross-Domain Audio Deepfake Detection (CD-ADD) dataset [18] was developed to support detection models across varying domains. A separate dataset utilizing retrieval-based voice conversion systems such as DEEP-VOICE [16] was also introduced. Furthermore, to support cross-lingual evaluation of detection systems, the DECRO dataset [6] was created, incorporating speech samples in both English and Mandarin.

Several datasets have subsequently been developed to support languages beyond English and Chinese. The WaveFake dataset [10] includes generated speech synthesized using neural vocoder models, covering both English and Japanese. A prominent multilingual dataset is MLAAD [17], which spans 38 languages and is built on the M-AILABS dataset [25], originally composed of recordings in eight languages sourced from audiobooks and interviews. The generated speech in MLAAD is synthesized using 82 TTS models across 33 architectures, including the Griffin-Lim vocoder. MLAAD has demonstrated strong utility by enabling the training of deepfake detection models that

TABLE 1. Comparison of existing datasets for deepfake speech detection, including language coverage, year, deepfake speech generation methods, number of utterances (all, pristine, and generated), and primary focus task addressed in each dataset.

Dataset	Year	Language(s)	Generation methods	# Utts.	# Pristine	# Generated	Focus task(s)
ASVspoof2015 [7]	2015	English	TTS, Vocoder, VC	263,151	16,651	246,500	Spoofing detection (speech synthesis and voice conversion)
ASVspoof2017 [8]	2017	English	Replay	18,030	3,565	14,465	Spoofing detection (replay attacks)
Fake or Real [9]	2019	English	TTS	>198,029	>110,744	87,285	General deepfake speech detection
ASVspoof2019 [2]	2019	English	TTS, Vocoder, VC, Replay	339,891	41,373	298,518	Spoofing detection (speech synthesis, voice conversion, and replay attacks)
WaveFake [10]	2021	English, Japanese	TTS, Vocoder	117,985	0*	117,985	General deepfake speech detection
ASVspoof2021 [11]	2021	English	TTS, Vocoder, VC, Replay, Hybrid	1,566,273	145,669	1,420,604	General deepfake speech and spoofing detection (speech synthesis, voice conversion, replay attacks)
ADD2022 [12]	2022	Mandarin	TTS, VC, Partially Fake	53,577	5,619	47,958	General deepfake speech detection
CFAD [13]	2022	Mandarin	TTS, Partially Fake	347,400	115,800	231,600	General deepfake speech detection (robustness, generalization)
In-the-Wild [14]	2022	English	TTS	31,779	19,963	11,816	Real-world deepfake speech detection
ADD2023 [15]	2023	Mandarin	TTS, VC, Partially Fake	517,068	243,194	273,874	General deepfake speech detection (include manipulation region location and deepfake algorithm recognition)
DEEP-VOICE [16]	2023	English	VC (retrieval-based)	7,484	3,742	3,742	Voice conversion deepfake detection
DECRO [6]	2023	English, Mandarin	TTS, VC	118,381	33,702	84,679	Cross-language deepfake speech detection
MLAAD [17]	2024	Multilingual (38 lang.)	TTS, Vocoder	154,000	0*	154,000	Multilingual deepfake speech detection
CD-ADD [18]	2024	English	TTS	145,570	25,111	120,459	Cross-domain deepfake speech detection
ASVspoof5 [19]	2024	English	TTS, Vocoder, VC, AT	>1,293,892	252,050	>1,041,842	General deepfake speech and spoofing detection (including adversarial attacks)
DFADD [20]	2024	English	TTS	207,955	44,455	163,500	Spoofing detection (diffusion and flow-matching based TTS)
CVoiceFake [21]	2024	Multilingual (5 lang.)	Vocoder	1,254,893	0*	1,254,893	Multilingual deepfake speech detection and speech content privacy preservation
SAFE Challenge [22]	2025	Multilingual	Unknown	Unknown	Unknown	Unknown	Unseen and various deepfake speech detection (robustness)
JMDS-Open (ours)	2025	Multilingual (15 lang.)	TTS, Vocoder, VC, AT	309,811	69,392	240,419	Multilingual deepfake speech detection
JMDS-All (ours)	2025	Multilingual (17 lang.)	TTS, Vocoder, VC, AT	353,845	77,759	276,086	Multilingual deepfake speech detection

* Pristine data are sourced from LJSpeech [23] and JSUT [24] (WaveFake); M-AILABS [25] (MLAAD); and CommonVoice [26] (CVoiceFake).

outperform those trained on other datasets such as In-the-Wild and Fake or Real. Its extensive linguistic diversity also facilitates robust cross-lingual evaluation and improves generalization. Another large-scale multilingual dataset is CVoiceFake [21], which contains over 1.25 million bonafide and deepfake utterances across five languages, with speech data sourced from the CommonVoice dataset [26].

Recent efforts have also focused on detecting deepfake speech in low-resource languages. Notably, studies have targeted languages within the ASEAN region [33], aiming to develop spoofing countermeasures tailored to these linguistic contexts. This line of work includes the construction of dedicated datasets for Thai (ThaiSpoof [34]), Indonesian (InaSpoof [35], [36]), Vietnamese (VSASV [37]), and Burmese (UCSYSpoof [38]). These studies underscore several challenges in building effective detection models for underrepresented languages, such as limited access to high-quality human speech data, inconsistencies in dataset quality across languages, and the rapid advancement of realistic spoofing techniques.

Table 1 compares the existing datasets for deepfake speech detection. Although these datasets have contributed significantly to the field, most are still limited in linguistic diversity and lack balanced representation across languages. JMDS fills a critical gap by providing a balanced, multi-source framework specifically designed for cross-language and cross-model generalization. The key differentiators between JMDS and the most recent benchmarks, MLAAD and CVoiceFake, are as follows:

- **Pristine-Generated Symmetry:** Unlike MLAAD, which contains 38 synthesized languages but only 8 human-spoken (pristine) languages, JMDS ensures every generated utterance is paired with authentic human speech across all included languages. This prevents models from "shortcut learning" where they might mistakenly identify neural machine translation artifacts in MLAAD as spoofing cues.
- **Source Diversity & Integrity:** Beyond simple aggregation, JMDS employs a strict protocol sourcing data from diverse acoustic environments to ensure more robust detection evaluation.
- **Generalization Benchmarking:** While CVoiceFake offers a higher volume of utterances, JMDS provides greater diversity in both language coverage and spoofing attack types.
- **Robustness to Resource Imbalance:** JMDS explicitly includes realistic scenarios where certain languages

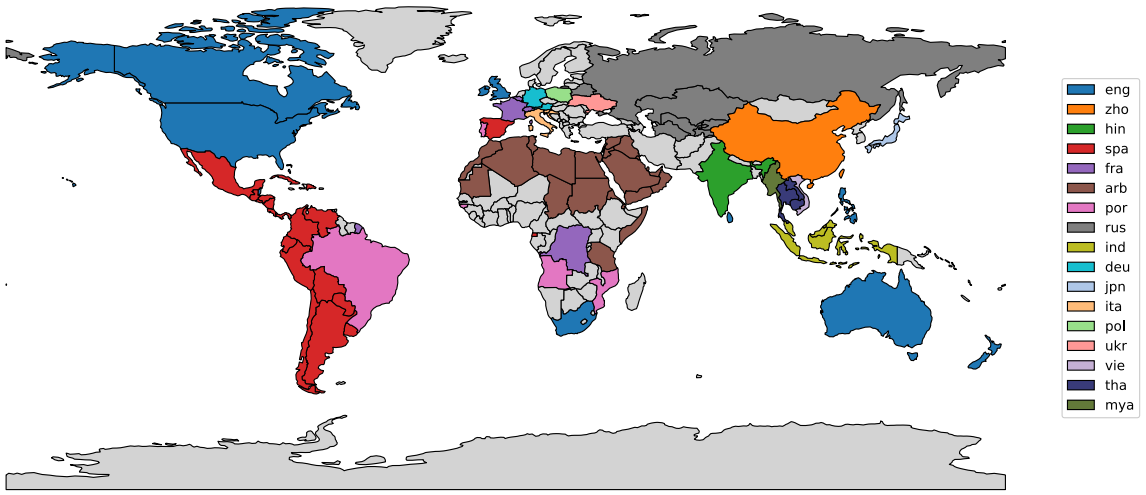

FIGURE 1. World map illustrating the primary geographical regions where the 17 languages included in our multilingual corpus are spoken as de facto and/or de jure languages. The legend provides the corresponding ISO 639 codes for each region.

(e.g., English) are high-resource while others have minimal pristine data.

III. JAIST MULTILINGUAL DEEPFAKE SPEECH (JMDS) DATASET

To facilitate the development of a reliable model for detecting generated speech across multiple languages, we assembled and carefully curated a diverse multilingual speech dataset, drawing from various open and private sources to ensure high quality and broad linguistic coverage.

A. CURATION PROCESS

To begin the dataset curation process, we first compiled a comprehensive list of publicly available speech corpora spanning various languages. From these, we selected 17 representative spoken languages: English, Mandarin, Hindi, Italian, Modern Standard Arabic, Spanish, Polish, German, French, Russian, Portuguese, Japanese, Ukrainian, Vietnamese, Thai, Indonesian, and Burmese. Figure 1 illustrates the geographical areas where these languages function as official or widely spoken primary languages.

To ensure high-fidelity human speech (pristine) and maintain relevance with current detection system technologies, we prioritized widely utilized public corpora, primarily developed for spoofing detection, speech recognition, or synthesis tasks. Table 2 provides a detailed breakdown of the dataset composition, including:

- the source repository name(s) and corresponding languages used in our dataset;
- the type of speech included (pristine, generated, or both), along with the associated deepfake or generation algorithms;
- a summary of the speech conditions, such as audio format and sampling rates.

For English, we utilized well-established corpora from the ASVspoof Challenge [28] for both pristine and generated speech, which comprise the largest portion of our dataset. We chose the latest challenge version to best align with current detection system development. This version builds on the MLS English data [29] and includes stronger attacks, featuring advanced text-to-speech (TTS), vocoder, voice conversion (VC), and adversarial attacks (AT) designed to mislead automatic speaker verification (ASV) and countermeasure (CM) systems. Codec compression was also applied to both pristine and generated speech to simulate realistic audio transmission conditions.

For Chinese, we adopted corpora from the Audio Deepfake Detection (ADD) Challenge [12] to supplement generated data. The pristine Chinese data was sourced from AISHELL-3 [31], which is the same corpus referenced by the ADD Challenge, ensuring the inclusion of high-quality human speech.

To support multilingual diversity in both pristine and generated data, we utilized established corpora such as M-AILABS [25] and MLAAD [17]. In addition, we augmented the pristine data for specific languages using the following resources: IndicVoice [40] for Hindi, MediaSpeech [39] for Arabic, CORAA [41] for Portuguese, JVS [42] for Japanese, Lotus [43] for Thai, and VIVOS [44] for Vietnamese.

In some cases, open-source corpora lacked sufficient quantity or quality of human speech. Where feasible, we supplemented the dataset with internally recorded speech. For example, we included English utterances spoken by non-native speakers from Asia to increase accent diversity. These hybrid sources are referred to as ''mixed-source'' repositories. The inclusion of these recordings helped ensure broad linguistic coverage and representativeness.

TABLE 2. Overview of data sources in the JMDS dataset, detailing the adopted languages in ISO 639 codes, data types (pristine and/or generated), deepfake methods (TTS: text-to-speech, vocoder, VC: voice conversion, AT: adversarial attack), audio sources (pristine only), audio conditions (C: Clean, M: Mixed, N: Noisy), file formats, and sample rates (Hz). "Mult." in the format and sample rate columns indicates that the original source recordings contain varying formats and/or sample rates.

Source	ISO Language(s)	Type	DF Methods	Audio Source (P)	Cond.	Fmt	SR
ASVspoof [19]	eng	P, G	TTS, Voc, VC, AT	LibriVox audiobooks	M	FLAC	16k
ADD [12]	zho	G	TTS, VC	-	N	WAV	16k
MLAAD [17]	hin, spa, fra, arb, tha, por, rus, deu, jpn, ita, pol, ukr, vie	G	TTS, Voc	-	N	WAV	22k
AISHELL-3 [31]	zho	P	-	Scripts, e.g. geographic news, smart home commands, number strings	C	WAV	16k
M-AILABS [25]	spa, fra, rus, deu, ita, pol, ukr	P	-	Audiobooks from LibriVox, Project Gutenberg, and Ukranian entities	C	WAV	16k
MediaSpeech [39]	arb	P	-	News broadcasting from Youtube	M	WAV	16k
Indic-voice [40]	hin	P	-	Read, extempore, and conversational (various tasks)	C, N	WAV	8k
CORAA [41]	por	P	-	Interviews, lectures, talks, dialogues, and reading tasks	C, N	WAV	16k
JVS [42]	jpn	P	-	Parallel and non-parallel sentences, incl. whispers and falsetto	C	WAV	24k
Lotus [43]	tha	P	-	Articles and text corpus (phonetically balanced)	C, M	WAV	16k
VIVOS [44]	vie	P	-	News and forums	C	WAV	16k
Private	mya	P, G	TTS, Voc, VC	Travel corpus (female speakers)	C	WAV	16k
Private	tha	P, G	TTS, Voc, VC	News (self-recorded) and Common Voice	C, M	WAV	Mult.
Private	ind	P, G	TTS, Voc, VC	Formal meetings, news, and audiobooks (Common Voice, Librivox, self-recorded)	C, N	Mult.	Mult.
Private	eng (nonnative)	P, G	TTS, Voc, VC	News (self-recorded)	C, N	WAV	Mult.

Note: Mixed (M) refers to varying non-studio acoustic environments (e.g., office, broadcast); Noisy (N) refers to explicit additive environmental interference.

Private-source data were also incorporated from prior studies on spoofing detection in Asian languages [33], [34], [35], [36], [38], [45], each containing pristine and generated speech synthesized from a variety of TTS, vocoder, and VC algorithms. The pristine recordings were typically collected in diverse environments using multiple devices, enriching the dataset's acoustic variability and thereby improving generalization for detection models.

We further observed that some repositories, such as MLAAD, included only a single speaker to generate spoofed data. Therefore, we limited the number of utterances to a maximum of 100 per spoofing algorithm in single-speaker settings, while retaining all utterances in multi-speaker settings from other speech corpora. Additionally, for datasets containing pristine speech from multiple speakers, such as M-AILABS, we carefully selected approximately 50 utterances per speaker. These efforts were made to ensure a balanced representation of speakers in each corpus.

After conducting the data acquisition from open-source, mixed-source, and private-source repositories, we organized the dataset into two configurations. The *Open Source* subset comprises only publicly available corpora (both open- and mixed-source) and serves as the primary contribution for promoting transparency and reproducibility. The *All Source* configuration additionally includes a limited amount of internally recorded speech (private-source), expanding language coverage and offering stronger baselines, particularly for languages with limited availability of high-quality public human speech data.

To reduce bias, we aimed for a balanced distribution of speech samples across different languages, setting the number of machine-generated utterances to be roughly four times that of human speech. This ratio reflects practical scenarios often encountered in deepfake speech detection. Additionally, given that many existing detection systems are primarily developed and tested using English data, we intentionally included a larger volume of English samples. An overview of the dataset composition across 17 languages is presented in Table 3.

During the final standardization phase, all audio clips were limited to a maximum duration of 60 seconds to maintain computational efficiency and temporal consistency across the training and evaluation pipelines. Each sample underwent a rigorous acoustic normalization process, involving

resampling to 16 kHz and conversion to a mono-channel format.

To ensure the integrity of the corpus, we implemented an automated filtering protocol to identify and remove malformed artifacts. This included verifying header compliance with the standard WAV format and performing a bit-stream check to exclude zero-byte files or corrupted segments that could trigger failures during feature extraction. Through this meticulous curation and validation process, we have established a high-fidelity, multilingual speech corpus that provides a consistent and robust foundation for advancing research in synthetic speech detection.

TABLE 3. Summary of the number of utterances (# utts.) distribution across 17 languages in the JMDS dataset.

Language	Open Source (# utts.)		All Source (# utts.)	
	Pristine	Generated	Pristine	Generated
English (eng)	49,977	174,484	49,977	174,484
Mandarin (zho)	4,410	24,642	4,410	24,642
Hindi (hin)	4,850	1,959	4,850	1,959
Italian (ita)	2,060	4,263	2,060	4,263
Arabic (arb)	2,505	2,957	2,505	2,957
Spanish (spa)	270	4,742	270	4,742
Polish (pol)	100	4,870	100	4,870
German (deu)	476	4,316	476	4,316
French (fra)	298	4,413	298	4,413
Russian (rus)	148	3,711	148	3,711
Portuguese (por)	1,000	3,011	1,000	3,011
Japanese (jpn)	1,000	2,978	1,000	2,978
Ukrainian (ukr)	298	2,237	298	2,237
Vietnamese (vie)	1,000	961	1,000	961
Thai (tha)	1,000	875	3,660	12,440
Indonesian (ind)	0	0	2,974	14,098
Burmese (mya)	0	0	2,733	10,004
Total	**69,392**	**240,419**	**77,759**	**276,086**
Subtotal	**309,811**		**353,845**	

B. METADATA

Our standardized speech corpus is organized into a unified dataset structure consisting of audio samples and their corresponding metadata. Each audio file is annotated with the following metadata fields:

- General information: Identifiers for the utterance, speaker, speaker gender, and a label indicating whether the sample is genuine human speech (labeled as 'pristine') or machine-generated (labeled as 'generated').
- Codec type: Specifies the audio codec applied to the sample before it is converted to WAV format, if any (e.g., M4A).
- Deepfake algorithm: Describes the method or model used to generate the synthetic or manipulated speech sample.

C. STATISTICS

Our compiled dataset comprises a total of 353,845 speech samples across 17 languages, including English and Chinese, which represent some of the world's most widely spoken languages.

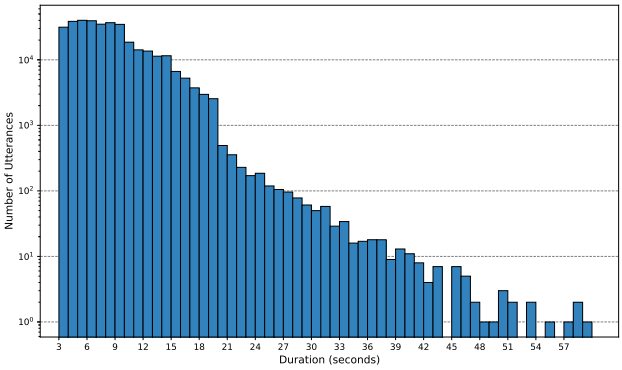

FIGURE 2. Histogram showing the distribution of utterance durations in the JMDS dataset (All Source) on a logarithmic scale.

To analyze the temporal characteristics of the dataset, we examined the distribution of utterance durations, as depicted in Fig. 2. Most samples fall within the range of approximately 3 to 10 seconds. The distributions are shown on a logarithmic scale to accommodate the wide variation in utterance counts across languages. Most languages exhibit a reasonable spread across duration categories, which supports generalization for models trained to detect speech artifacts across diverse speaking styles and utterance lengths, thereby enhancing the robustness of evaluation systems under real-world conditions.

Our compiled dataset encompasses a multitude of deepfake methods for speech generation, categorized into four distinct types: TTS-based attacks, vocoder-based attacks, VC-based attacks, and adversarial attacks (AT). TTS-based attacks are generated using text-to-speech systems that synthesize speech directly from textual input, often leveraging models such as GlowTTS, VITS, Tacotron2, or other pretrained neural architectures. Vocoder-based attacks generate speech from intermediate acoustic features such as mel spectrograms, using vocoders like Griffin-Lim. VC-based attacks modify a source speaker's voice to resemble that of a target speaker, typically without changing the linguistic content, using models such as StarGANv2-VC or ASR-based VC pipelines. AT introduces subtle, imperceptible perturbations to utterances, specifically optimized to degrade the performance of spoofing detection systems [19].

We incorporated generated speech samples from four different data sources: ASVspoof [28], ADD [12], MLAAD [17], and a portion of our private-source data. Among all data sources included in the JMDS dataset, only ASVspoof explicitly incorporates adversarial attacks as part of its deepfake generation strategy. In this case, adversarial perturbations are applied as a post-processing step to spoofed utterances produced by TTS, vocoder, or VC systems. Portions of both pristine and generated samples have also been subjected to encoding and compression using speech

codecs, simulating real-world conditions where speech may be transmitted over networks or stored in compressed formats.

For the MLAAD set, we selected ten speech generation algorithms, along with their variants and modifications, including TTS models and vocoders tailored to the adopted languages in the JMDS dataset. The ADD dataset also

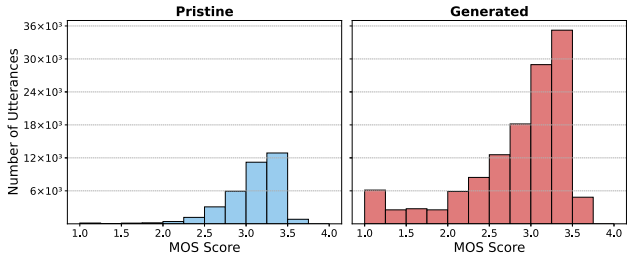

FIGURE 3. Distribution of MOS scores in the JMDS dataset (All Source).

contains generated samples produced utilizing commonly used TTS and VC algorithms, although the exact models are not disclosed. We used the clean subset of ADD to construct the training and development data. Our private-source generated speech was created using high-performance speech synthesis systems, encompassing TTS, vocoders, VC, and several proprietary TTS models. The term "system" was chosen to reflect the inclusive scope of the deepfake generation methods, encompassing models (e.g., VITS, Tacotron2), techniques or components (e.g., Griffin-Lim vocoder, Malafide attack), and frameworks or pipelines (e.g., Whisper-based TTS, unit-selection-based TTS).

D. AUDIO QUALITY

We analyze the quality of the audio data using the Mean Opinion Score (MOS) obtained from DNSMOS [46], a robust and non-intrusive perceptual objective speech quality metric. DNSMOS is well-suited for measuring audio quality as it serves as a proxy for non-intrusive objective evaluation. The MOS scale ranges from very poor (MOS = 1) to excellent (MOS = 5). We evaluated both pristine and generated speech across multiple languages and data sources to provide a comprehensive understanding of the dataset's perceptual quality.

Figure 3 shows the distribution of MOS scores for pristine (left) and generated (right) speech in the *All Source* configuration of our dataset. Pristine speech samples generally exhibit moderately high MOS scores, with an average of 3.14, indicating good overall perceptual quality. Only a few utterances are scored below 2.5, suggesting that most human recordings are relatively clean and clear. While many generated utterances achieve MOS scores comparable to those of pristine samples, particularly in the 3.0 to 3.5 range, a notable portion falls below 2.5. This drop indicates inconsistencies in quality introduced by certain generation algorithms. Moreover, the broader spread and higher variance in MOS scores for generated speech highlight the varying quality among synthetic samples.

Although this variability in audio quality may help promote generalizability in spoof detection systems, extremely low-quality synthetic speech could introduce undesirable noise during model training. Nevertheless, we feel that retaining such low-quality samples in the evaluation set is beneficial, as it ensures the inclusion of various audio conditions and contributes to a more robust final evaluation.

IV. SPOOF DETECTION METHODS

We utilized several state-of-the-art models to evaluate the proposed dataset. This section outlines the experimental setup, including preprocessing strategies and detailed parameter configurations for each model. We primarily experimented with two model families: Residual Network (ResNet)-based [47] and AASIST-based [48] architectures. We chose the ResNet family as a robust, historically validated baseline from the ASVspoof challenges, valued for its ability to learn complex spectro-temporal patterns from handcrafted features. Conversely, the AASIST architecture represents the current state-of-the-art, chosen for its end-to-end capability to learn low-level artifacts directly from raw waveforms, and its superior modeling of long-term dependencies via attention mechanisms. Together, these two architectures allow us to compare conventional feature-based methods with modern raw waveform-based approaches for anti-spoofing.

To address the issue of varying audio sample lengths within the dataset, we implemented a preprocessing step to standardize the input duration. Shorter audio clips were padded by repeating their content until the target length was achieved, while longer clips were truncated. Our initial target duration was 4 seconds. However, recognizing that forged speech samples might require more extensive temporal information for accurate classification and to mitigate potential misclassifications due to limited feature representation, we subsequently increased the target durations to 10 seconds. This extension aimed to preserve richer contextual information and ultimately enhance the classification accuracy.

A. CQT-ResNet34

For our experiments, we selected the ResNet34 model [47] and used Constant-Q Transform (CQT) features [49] as input. The CQT was chosen because it provides a superior time-frequency representation compared to the Short-Term Fourier Transform (STFT). By maintaining a constant Q factor, CQT achieves better temporal resolution at higher frequencies and better frequency resolution at lower frequencies. This capability is vital for capturing the subtle acoustic characteristics needed to distinguish between genuine and deepfake speech, aligning with its proven efficacy in related fields like acoustic scene classification. We tested the model using both 4-second and 10-second speech segments.

B. RawSpeech-AASIST

We adopted the AASIST architecture for its end-to-end capability to extract relevant features directly from the raw waveform inputs [48]. This motivated our use of raw waveforms

for the RawSpeech-AASIST model. We noted an empirical relationship where extending the input segment length led to improved spoofing detection performance. However, this gain incurred a significant increase in computational cost. To balance the performance gains against computational efficiency, we benchmarked the model's performance using two distinct input speech durations: 4 seconds and 10 seconds.

C. SSL-AASIST

We further investigated a variation of AASIST that integrates self-supervised learning (SSL) features. In this approach, we leveraged pre-trained SSL models to extract rich representations from the raw audio, which were subsequently utilized as input to the AASIST architecture. Our experiments explored two different input lengths: 4 seconds and 10 seconds. For SSL feature extraction, we selected two high-performing pre-trained models known for their effectiveness across various speech tasks: XLS-R with 300 million parameters **XLS-R 300M** and **WavLM-Large**. These models were chosen for their ability to capture informative acoustic patterns.

D. RawSpeech-RawNet3

RawNet3 represents a significant leap in end-to-end deepfake detection by bypassing handcrafted features and operating directly on raw audio waveforms [50]. By integrating parameterized analytic filterbanks and Res2Net-based backbone blocks, the model excels at capturing fine-grained temporal artifacts and high-frequency inconsistencies—often exceeding 4000 Hz—where synthetic speech generators typically leave "digital fingerprints."

Furthermore, RawNet3 achieved state-of-the-art performance in benchmark speaker verification tasks on the VoxCeleb dataset. When paired with robust training strategies, such as Frequency-Selective Adversarial Training (F-SAT) or self-supervised pre-training, RawNet3 demonstrates a superior capacity for open-world generalization. This makes it a formidable tool for identifying hyper-realistic voice clones that traditional spectrogram-based models might overlook.

V. GENERAL EVALUATION

A. EVALUATION METRICS

Given the straightforward nature of generated speech detection as a binary classification task—distinguishing between pristine (positive) and generated (negative) speech—we selected balanced accuracy as our evaluation metric. Balanced accuracy, calculated as the average of the accuracy in the pristine class and the accuracy in the generated class, provides a more robust measure of performance, especially in potentially imbalanced datasets. The calculation for balanced accuracy is as follows.

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

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

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

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 deepfake speech detection [51]. 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.

TABLE 4. Distribution of the JMDS dataset into training (Train), development (Dev), and evaluation (Eval) sets, as used for model validation during our experiments.

Partition	Subset	# utts.		
		Pristine	Generated	Total
Open	Train	28,981	100,452	129,433
	Dev	18,269	73,050	91,319
	Eval	18,296	73,098	91,394
All	Train	44,442	137,723	182,165
	Dev	24,977	90,852	115,829
	Eval	24,547	89,478	114,025

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 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.

To ensure a robust and unbiased evaluation across the diverse linguistic categories in our study, we adopted a stratified sampling strategy and macro-averaged metrics. Specifically, the stratified sampling was implemented at the speaker level to ensure that the ratio of pristine to generated samples remained consistent across the training, development, and evaluation splits, while preventing speaker leakage. To further mitigate the statistical dominance of

TABLE 5. Cross-partition evaluation on the JMDS dataset, comparing models trained on the JMDS-Open subset versus the JMDS-All subset. All training configurations utilized a 4-second audio padding.

			Evaluation data →	JMDS-Open				JMDS-All			
Front-end	Back-end	Training data ↓	Label	Accuracy (%)	Balanced Acc. (%)	AUC (%)	EER (%)	Accuracy (%)	Balanced Acc. (%)	AUC (%)	EER (%)
CQT	ResNet	JMDS-Open	Pristine	35.01	53.26	59.73	43.29	34.61	51.59	58.52	43.56
			Generated	71.52				68.57			
RawSpeech	AASIST		Pristine	70.51	79.52	91.42	14.63	66.64	76.50	90.29	16.66
			Generated	88.52				86.37			
RawSpeech	RawNet3		Pristine	67.79	77.62	89.98	15.78	65.95	76.17	88.89	16.83
			Generated	87.44				86.39			
XLS-R (300M)	AASIST		Pristine	76.03	83.79	94.57	11.00	69.78	79.01	92.72	14.46
			Generated	91.55				88.24			
CQT	ResNet	JMDS-All	Pristine	36.76	52.11	61.37	41.89	36.37	50.63	60.76	42.44
			Generated	67.45				64.90			
RawSpeech	AASIST		Pristine	73.87	82.05	94.24	11.83	75.50	83.36	94.74	11.01
			Generated	90.23				91.23			
RawSpeech	RawNet3		Pristine	69.07	78.63	90.05	15.10	69.55	79.04	90.69	14.41
			Generated	88.20				88.54			
XLS-R (300M)	AASIST		Pristine	81.09	87.48	96.19	8.32	80.21	86.89	96.29	8.46
			Generated	93.87				93.56			

the English subset and ensure that low-resource languages were equally represented in our final analysis, we utilized macro-averaged EER (Macro-EER) and Macro-Accuracy. Unlike micro-averaging, which pools all samples together and can be skewed by the majority class, macro-averaging treats each language as an independent entity, providing a more equitable assessment of the system's cross-lingual generalization capabilities.

B. DATA PARTITION

Table 4 summarizes the distribution of the training, development, and evaluation sets across the partitioned data. To validate our dataset design and collection process, we conducted experiments aimed at developing a robust and generalized spoofing detection system.

We partitioned the dataset into three mutually exclusive subsets—training, development, and evaluation—following a 6:2:2 ratio. The split was primarily determined by speaker ID and source dataset. Additionally, where metadata was available, we balanced the distribution of attack IDs and gender to ensure each subset maintained consistent attribute characteristics. To maintain experimental integrity, there is no overlap of speech samples between the subsets.

C. RESULTS AND ANALYSIS

Combining multiple data sources with varying quality can be challenging and may inadvertently degrade the performance of a trained model. To address this, we conducted a thorough cross-evaluation using the two partitions of our dataset: one comprising only open-source data (Open) and the other incorporating private data sources (All). Specifically, we trained representative methods on the Open partition (JMDS-Open) and evaluated them on both the Open and All partitions. Conversely, we also trained on the All partition (JMDS-All) and evaluated on both. The results of this cross-evaluation are presented in Table 5. All model architectures were trained utilizing a fixed-length audio padding approach implemented via repeat padding. For samples exceeding the target duration, a center-truncation strategy was applied to ensure temporal consistency across the batch.

In our initial cross-partition evaluation on the JMDS dataset, we observed significant performance disparities across various model architectures. The classical approach, utilizing Constant Q Transform (CQT) features with a ResNet34 backbone, struggled to accurately detect pristine signals and frequently misclassified all speech as generated, resulting in low balanced accuracy and a high Equal Error Rate (EER). In contrast, the end-to-end AASIST architecture demonstrated superior performance using raw waveforms (RawSpeech). When trained and evaluated on JMDS-Open, AASIST achieved a balanced accuracy of 79.52%, an AUC of 91.42%, and an EER of 14.63%. Although RawNet3 is often cited for its efficacy with raw speech, it consistently underperformed relative to AASIST by approximately 1–2%, suggesting weaker generalization capabilities in this context.

The integration of Self-Supervised Learning (SSL) features, particularly XLS-R (300M), further enhanced these results, yielding a balanced accuracy of 83.79%, an AUC of 94.57%, and a reduced EER of 11.00%. These improvements indicate that high-capacity pre-trained models significantly bolster the system's ability to distinguish between genuine and spoofed speech. Overall, these performance trends remained consistent across various front-end and back-end pairings, confirming that the combination of SSL features and the AASIST framework provides the most robust solution for the JMDS dataset.

Furthermore, our analysis of the Open and All partitions revealed that training on JMDS-All generally yielded superior performance compared to training on JMDS-Open. This performance boost is likely attributable to the inclusion of controlled-environment recordings and low-resource language data within the private portion of the "All" partition. By introducing these variables, the dataset increases the overall complexity and difficulty of the detection task. Consequently, the JMDS-All partition serves as a more rigorous and comprehensive benchmark for evaluating the robustness of deepfake speech detection models.

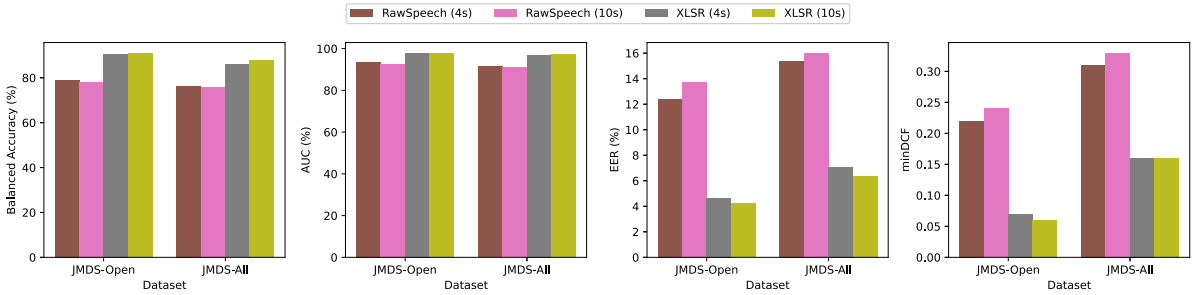

FIGURE 4. Performance comparison of AASIST-based models trained on JMDS-Open using 4s and 10s padding, evaluated on JMDS-Open and JMDS-All datasets.

TABLE 6. Comparison on various pair of front-end and back-end methods using 10-second padding and evaluation on JMDS-Open evaluation data.

Front-end	Back-end	Label	Accuracy (↑) (%)	Balanced (↑) Acc. (%)	AUC (%) (↑)	EER (%) (↓)	minDCF (↓)
CQT	ResNet	Pristine	35.01	53.26	59.73	43.29	1.00
		Generated	71.52				
XLS-R (300M)	ResNet	Pristine	43.76	68.88	84.96	22.65	0.58
		Generated	94.01				
RawSpeech	AASIST	Pristine	71.84	80.45	91.22	14.79	0.24
		Generated	89.07				
RawSpeech	RawNet3	Pristine	71.50	80.42	91.24	14.38	0.28
		Generated	89.33				
WavLM	AASIST	Pristine	75.15	83.27	95.24	10.97	0.26
		Generated	91.39				
XLS-R (300M)	AASIST	Pristine	**80.88**	**87.29**	**96.10**	**8.22**	**0.16**
		Generated	**93.70**				

TABLE 7. Cross-dataset evaluation results for various front-end/back-end method pairings. Training data: ASVspoof2019, ASVspoof2024, JMDS-Open. Evaluation data: FoR, ASVspoof2019, in-the-wild (ITW), DECRO, JMDS-Open, SAFE challenge 2025 (Task 1) (SAFE). For representative evaluation metrics, we utilized equal error rate (EER) (↓) and balanced accuracy (Acc.) (↑) in percentage (%).

		Eval. Data →	FoR [9]		ASVspoof2019 [2]		ITW [14]		DECRO [6]		JMDS-Open		SAFE [22]
Front-end	Back-end	Train. Data ↓	EER (%)	Acc. (%)	EER (%)	Acc. (%)	EER (%)	Acc. (%)	EER (%)	Acc. (%)	EER (%)	Acc. (%)	Acc. (%)
CQT	ResNet	ASVspoof2019	46.85	65.93	14.96	67.40	47.97	53.47	23.85	71.76	44.97	53.74	48.58
RawSpeech	AASIST		12.51	88.13	2.42	93.73	39.59	63.02	35.18	65.74	35.96	59.25	42.70
CQT	ResNet	ASVspoof2024	33.39	65.36	23.82	60.98	36.34	64.17	34.50	59.67	44.92	56.76	36.73
RawSpeech	AASIST		7.15	92.39	21.76	63.20	16.39	83.62	35.77	73.70	23.15	68.71	47.85
CQT	ResNet	JMDS-Open	18.44	75.20	23.52	59.78	45.72	53.28	29.15	65.64	43.29	53.26	58.74
RawSpeech	AASIST		**0.05**	**99.87**	14.43	**71.62**	**6.45**	**93.51**	**15.95**	**82.11**	14.79	80.45	**67.27**
XLS-R	AASIST		11.71	81.75	**13.90**	70.75	11.66	88.52	30.58	64.26	**8.22**	**87.29**	60.73

In our subsequent analysis, we investigated the impact of various padding lengths on models built using the AASIST architecture. Our experiments indicate that the chosen padding duration is often critical for accurately detecting artifacts in generated speech, though its importance varies depending on the characteristics of the evaluation set. For instance, while shorter padding may suffice for datasets primarily consisting of brief utterances (under 5 seconds), the diverse utterance lengths within our JMDS dataset necessitated a comparison between 4-second and 10-second padding configurations.

Figure 4 illustrates the performance of these models—evaluated via balanced accuracy, AUC, EER, and minDCF—on both JMDS-Open and JMDS-All, with all models trained on JMDS-Open. When utilizing raw waveforms as input, the performance disparity between the two padding lengths was minimal. However, we observed a slight improvement when integrating XLS-R features with the longer 10-second padding. To further assess generalization on entirely unseen data, we evaluated these models on the SAFE Challenge 2025 [22], [52]. On this external dataset, which includes audio samples spanning up to 60 seconds, increasing the padding size yielded a significant performance boost—exceeding 5% in balanced accuracy. A more comprehensive analysis of this cross-dataset evaluation is provided in Section VI-B.

To ensure a fair comparison across various front-end and back-end configurations, all models were trained and evaluated on the JMDS-Open dataset using a 10-second padding duration. The results, summarized in Table 7, compare six specific combinations: (1) CQT-ResNet, (2) XLSR-ResNet, (3) RawSpeech-AASIST, (4) RawSpeech-RawNet3, (5) WavLM-AASIST, and (6) XLSR-AASIST.

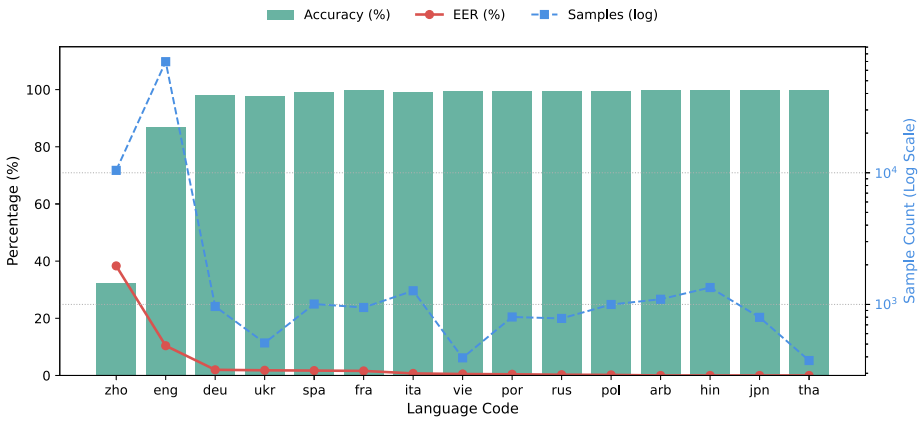

FIGURE 5. Language-specific detection performance of the RawSpeech-AASIST model on the JMDS-Open dataset. The results illustrate the model's varying robustness across 15 languages.

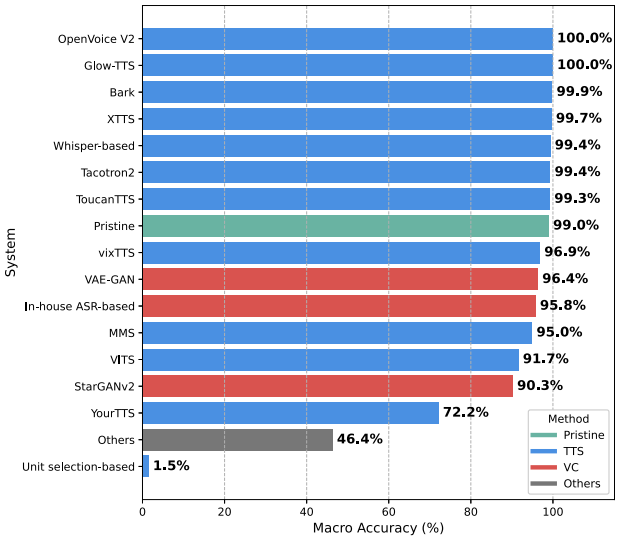

FIGURE 6. Attack-specific detection accuracy of the RawSpeech-AASIST model on the JMDS-Open dataset.

Each configuration was assessed using the full suite of evaluation metrics detailed in Subsection V-A.

Our initial comparison between traditional CQT features and those extracted from the pre-trained XLS-R (300M) SSL model—both utilizing a ResNet backbone—revealed a substantial performance gap. The pre-trained SSL features yielded significantly more robust results, albeit at a higher computational cost. Furthermore, AASIST-based architectures generally outperformed their ResNet counterparts; for instance, using XLSR as a front-end, AASIST achieved a balanced accuracy of approximately 87.29%, compared to 68.88% for the ResNet model. Finally, our findings emphasize the importance of selecting a domain-appropriate SSL model. Although WavLM frequently outperforms XLS-R in standard speech processing benchmarks [53], XLS-R demonstrated superior efficacy in this deepfake detection task. Specifically, XLS-R achieved over 4% higher balanced accuracy, a 2% lower EER, and a 0.1 lower minDCF compared to the WavLM-based front-end.

VI. MODEL ROBUSTNESS AND GENERALIZATION EVALUATION

This section evaluates model performance across the various languages, attack types, and speech quality represented in the JMDS dataset. Furthermore, a cross-dataset evaluation is conducted to assess generalization capabilities on external benchmarks.

A. ATTRIBUTE-SPECIFIC: LANGUAGE AND ATTACK TYPE

Figure 5 shows the language-specific performance of the AASIST model using RawSpeech as input. Model performance varies significantly across the 15 languages in the JMDS-Open dataset, with results heavily influenced by two distinct outliers: Mandarin Chinese (zho) and English (eng). While the architecture achieves near-perfect detection (0.0% EER) for low-resource languages such as Arabic, Hindi, and Japanese, it encounters a catastrophic failure in Mandarin, exhibiting an EER of 38.33% and a balanced accuracy of only 32.39%. This performance gap suggests that the synthesis artifacts in specific benchmarks, such as the Audio Deepfake Database (ADD) used for Mandarin, are uniquely subtle and acoustically distinct from the spectral cues learned during training.

The dual-axis analysis further highlights a "scaling parado" within the dataset distribution. In languages with fewer than 1,500 samples, the model generalizes exceptionally well, likely because these subsets utilize synthesis methods with easily identifiable vocoder artifacts. Conversely, in the high-resource English subset, the EER rises to 10.41%. This indicates that larger, more diverse subsets contain a broader range of sophisticated generation techniques that create significant overlap between genuine and spoofed speech in the feature space.

Ultimately, these findings confirm that detection difficulty is primarily dictated by synthesis sophistication rather than

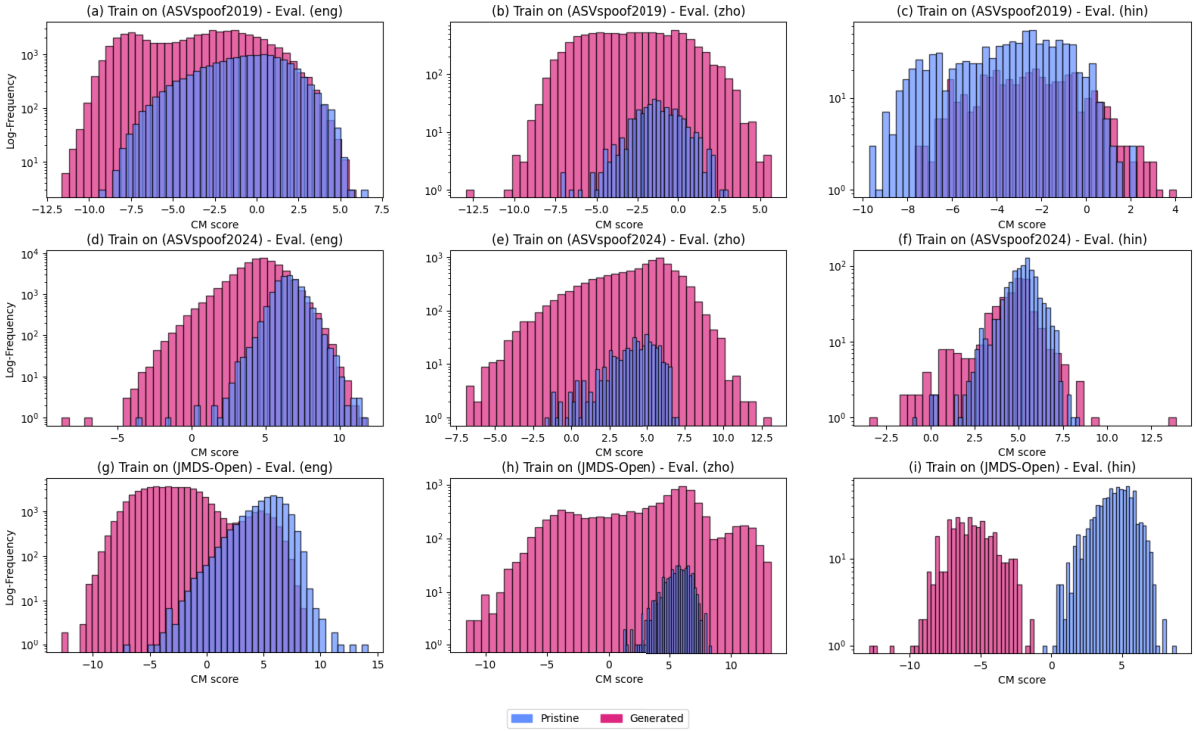

FIGURE 7. Language-specific performance using RawSpeech-AASIST method. Three most represented languages within JMDS-Open were included: English (eng), Mandarin Chinese (zho), and Hindi (hin).

linguistic properties. The high accuracy in low-resource languages likely reflects the presence of less advanced generation techniques, whereas the high error rates in English and Mandarin underscore the challenge of detecting high-fidelity deepfakes. Consequently, achieving a truly language-agnostic countermeasure requires exposure to a balanced variety of advanced attacks across all target languages to mitigate latent linguistic and algorithmic biases.

Figure 6 shows the attack-specific performance of the RawSpeech-AASIST model evaluated on JMDS-Open dataset. The robustness of the detection framework varies significantly depending on the underlying synthesis architecture, revealing that synthesis sophistication—rather than the broad category of TTS or VC—is the primary determinant of detection difficulty. A critical finding is the model's near-total failure to detect Unit selection-based TTS, which yielded an accuracy of only 1.54%. While neural models produce spectral artifacts that the AASIST front-end is optimized to identify, concatenative unit-selection methods rely on natural speech segments. This preserves local acoustic naturalness in a way that bypasses learned detection features, effectively allowing the synthetic signal to "hide" within the feature space of pristine speech.

In contrast, most neural-based TTS architectures, such as Glow-TTS, OpenVoice V2, and Bark, are detected with near-perfect accuracy. This suggests these models share common "fingerprints"—likely tied to standard vocoding stages like HiFi-GAN—that the countermeasure has successfully generalized. However, performance degrades against more advanced neural approaches like StarGANv2 (90.3%) and YourTTS (72.2%). Because VC models modify an existing human source signal, they retain more authentic prosodic and phase information than standard TTS.

Despite these challenges, the model maintains a high pristine accuracy of 99.04%, proving that high detection rates are not the result of excessive false positives. This stability confirms that the system has identified specific synthetic indicators rather than becoming over-sensitive to general speech variations. However, the significantly lower accuracy on the "Others" category (46.4%) highlights the risk of unknown attacks. These results underscore the necessity for continuous benchmark updates to include diverse, high-fidelity generative methods that currently test the limits of cross-dataset generalization.

B. CROSS-DATASET EVALUATION

To assess the generalization capabilities of the models beyond the JMDS dataset, a comprehensive cross-dataset evaluation was conducted. This involved training the architectures on specific benchmark datasets, namely ASVspoof 2019 [2], ASVspoof 2024 [51], and JMDS-Open. Performance was subsequently evaluated on these training sets as well as on entirely unseen datasets, including FoR [9], In-the-Wild (ITW) [14], DECRO [6], and the SAFE Challenge 2025 [22].

Table 7 presents the results of this cross-dataset analysis. Direct evaluation on the ASVspoof 2024 evaluation set

TABLE 8. Audio quality of the JMDS dataset measured using mean opinion score (MOS) for pristine and generated speech, along with its standard deviation.

Language	Pristine	Generated
Arabic (arb)	2.88 ± 0.33	3.29 ± 0.31
German (deu)	3.12 ± 0.31	3.09 ± 0.42
English (eng)	3.18 ± 0.27	2.69 ± 0.73
French (fra)	3.26 ± 0.21	3.18 ± 0.38
Hindi (hin)	2.81 ± 0.29	3.14 ± 0.40
Indonesian (ind)	2.95 ± 0.40	3.15 ± 0.36
Italian (ita)	3.17 ± 0.24	3.04 ± 0.47
Japanese (jpn)	3.18 ± 0.45	3.22 ± 0.45
Burmese (mya)	3.24 ± 0.19	3.13 ± 0.27
Polish (pol)	3.24 ± 0.24	3.15 ± 0.42
Portuguese (por)	2.42 ± 0.67	3.22 ± 0.40
Russian (rus)	3.29 ± 0.19	3.16 ± 0.41
Spanish (spa)	3.14 ± 0.28	3.15 ± 0.42
Thai (tha)	3.18 ± 0.27	2.97 ± 0.54
Ukrainian (ukr)	3.22 ± 0.22	2.92 ± 0.52
Vietnamese (vie)	3.22 ± 0.16	3.30 ± 0.21
Mandarin (zho)	3.19 ± 0.20	2.57 ± 0.53
All	$\mathbf{3.08 \pm 0.35}$	$\mathbf{2.84 \pm 0.64}$

was not possible due to the absence of public ground-truth labels. Instead, the ASVspoof 2024 training data was utilized to develop the models, which were then assessed on the remaining datasets. Regarding the SAFE benchmark, only the balanced accuracy is reported, as it is the sole metric provided on the challenge's official leaderboard.

The results presented herein were obtained using general model architectures and standard hyperparameter tuning on validation sets, without dataset-specific optimizations. The significant variability in balanced accuracy underscores the persistent challenge of generalization in this field, particularly when comparing the JMDS-Open dataset to existing benchmarks. While models demonstrate high efficacy when training and evaluation sets are aligned, performance diminishes considerably when applied to unseen datasets.

Consistent with prior experiments, AASIST-based architectures generally outperformed ResNet across all evaluated partitions. However, results indicate that in challenging, real-world environments involving entirely unseen data, detection performance remains a critical area for improvement. This is exemplified by the SAFE benchmark, where balanced accuracy reached only approximately 67%, highlighting a persistent need for robust cross-domain strategies to bridge the gap between training and real-world application.

The primary obstacle to improving robustness and generalization in deepfake speech detection is the challenge of evaluation across unknown attack vectors. While cross-lingual evaluation remains significant, models typically exhibit a more catastrophic performance drop when encountering unseen, high-quality generation methods—referred to as zero-shot attacks—than when processing an unseen language generated by a known method.

This trend is clearly supported by the results in Table 7. The SAFE challenge utilizes advanced, unknown attacks [22], leading to significantly lower detection accuracy. Conversely, the English-language FoR and ITW datasets, despite containing unknown attacks, are substantially easier to detect (accuracies of RawSpeech-AASIST typically exceeding 90%) because their underlying synthesis methods are less sophisticated. This contrast indicates that the sophistication of the generation method, rather than linguistic variation, is the primary factor limiting detection accuracy when a reliable training benchmark is available.

Furthermore, the JMDS-Open dataset serves as a more effective training source for generalization than the English-centric ASVspoof 2019 and ASVspoof 2024 datasets. For example, the top-performing model trained on JMDS-Open achieved a balanced accuracy of 71.62% on ASVspoof 2019 and 67.27% on the SAFE challenge. In contrast, models trained on either ASVspoof dataset performed significantly worse on the SAFE challenge, with balanced accuracies falling below 50%. This disparity demonstrates the superior cross-dataset generalizability of models trained on JMDS-Open.

To further investigate the language-specific performance of the top-performing architecture (RawSpeech-AASIST), an analysis was conducted on the three most prevalent languages in the JMDS-Open dataset: English (eng), Mandarin Chinese (zho), and Hindi (hin). The models for this evaluation were trained on a combined corpus of ASVspoof 2019, ASVspoof 2024, and the complete JMDS-Open dataset. Detailed results are illustrated in Fig. 7.

The language-specific evaluation revealed notable patterns in model generalization. As anticipated, performance was most robust when utilizing the model trained on the JMDS-Open subset. However, detection accuracy for Mandarin Chinese (zho) was consistently poor across all configurations, as shown in Fig. 7. While countermeasure scores for pristine Mandarin speech—evaluated via models trained on ASVspoof 2024 and JMDS-Open—showed a positive bias, the score distribution for generated Mandarin speech exhibited significant overlap with pristine samples. This diminished discriminative power in Mandarin may be attributed to the lower synthesis quality inherent in the Audio Deepfake Database (ADD) benchmark, where the Mean Opinion Score (MOS) typically falls below 3.0. These results highlight a critical requirement for enhanced multilingual robustness in future deepfake detection frameworks.

TABLE 9. Deepfake methods and systems used to generate samples in the JMDS dataset, along with their corresponding data sources. The ADD dataset is not included due to the undisclosed details of its generation methods.

Deepfake method	Systems	Link	Source
TTS	Bark	https://github.com/suno-ai/bark	MLAAD, Private-source
	XTTS	https://huggingface.co/coqui/XTTS-v2	MLAAD, ASVspoof
	MMS-TTS	https://huggingface.co/facebook/mms-tts	MLAAD, Private-source
	VITS	https://github.com/jaywalnut310/vits	MLAAD, ASVspoof
	Tacotron2	https://github.com/NVIDIA/tacotron2	MLAAD
	OpenVoice V2	https://huggingface.co/myshell-ai/OpenVoiceV2	MLAAD
	Glow-TTS	https://github.com/CODEJIN/Glow_TTS	MLAAD, ASVspoof
	Whisper-based TTS	https://github.com/WhisperSpeech/WhisperSpeech	MLAAD
	vixTTS	https://huggingface.co/capleaf/viXTTS	MLAAD
	Grad-TTS	https://grad-tts.github.io/	ASVspoof
	FastPitch	https://fastpitch.github.io/	ASVspoof
	ToucanTTS	https://toucantts.com/	ASVspoof
	YourTTS	https://github.com/Edresson/YourTTS	ASVspoof
	ZMM-TTS	https://github.com/nii-yamagishilab/ZMM-TTS	ASVspoof
	Unit selection-based	https://github.com/marytts/marytts	ASVspoof
	Proprietary TTS	–	Private-source
Vocoder	Griffin-Lim	https://github.com/emotechlab/griffin-lim	MLAAD
	WORLD	https://github.com/mmorise/World	Private-source
	Hifi-GAN	https://github.com/jik876/hifi-gan	Private-source
VC	StarGANv2-VC	https://github.com/yl4579/StarGANv2-VC	ASVspoof
	VAE-GAN	https://github.com/rishabhd786/VAE-GAN-PYTORCH	ASVspoof
	In-house ASR-based VC	–	ASVspoof
	DiffVC	https://github.com/trinhtuanvubk/Diff-VC	ASVspoof
	FreeVC	https://github.com/OlaWod/FreeVC	Private-source
	Spectral filtering	–	ASVspoof
Adversarial Attack	Malafide	https://github.com/eurecom-asp/malafide	ASVspoof
	Malacopula	https://github.com/eurecom-asp/malacopula	ASVspoof

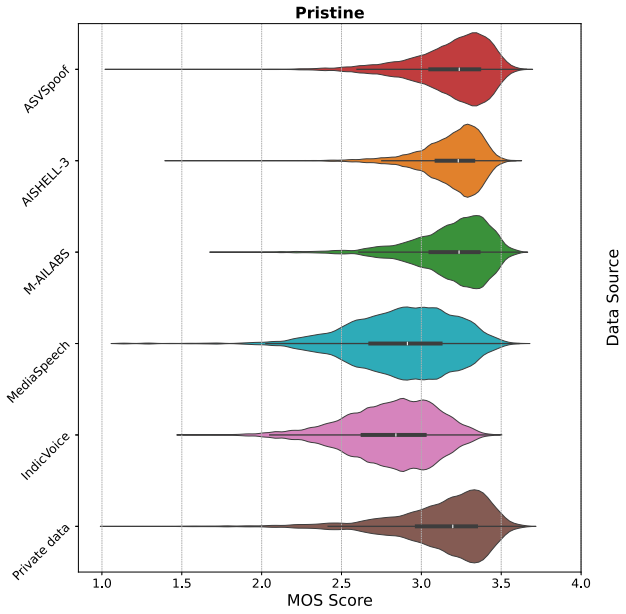

FIGURE 8. Violin plots illustrating the Distribution of MOS scores for pristine speech across different data sources in the JMDS dataset. Pristine sources generally show consistently high perceptual quality.

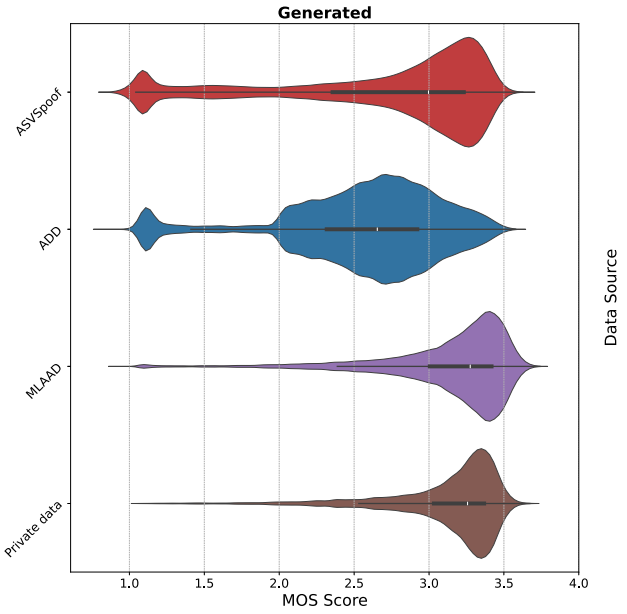

FIGURE 9. Violin plots illustrating the Distribution of MOS scores for generated speech across different data sources in the JMDS dataset. Among the generated sources, only ADD exhibits a notably broad range of audio quality, indicating varied synthesis characteristics, while others like ASVspoof, MLAAD, and Private data tend to produce more consistent high-MOS outputs.

VII. DISCUSSION

This study introduces the JMDS dataset, a comprehensive multilingual resource for synthetic speech detection, and evaluates its utility through benchmarking and cross-dataset experimentation. Our findings highlight several key aspects.

1) First, cross-evaluation within JMDS revealed that the partitions containing private data, while inherently more challenging, provide a more rigorous and realistic testbed for model robustness.

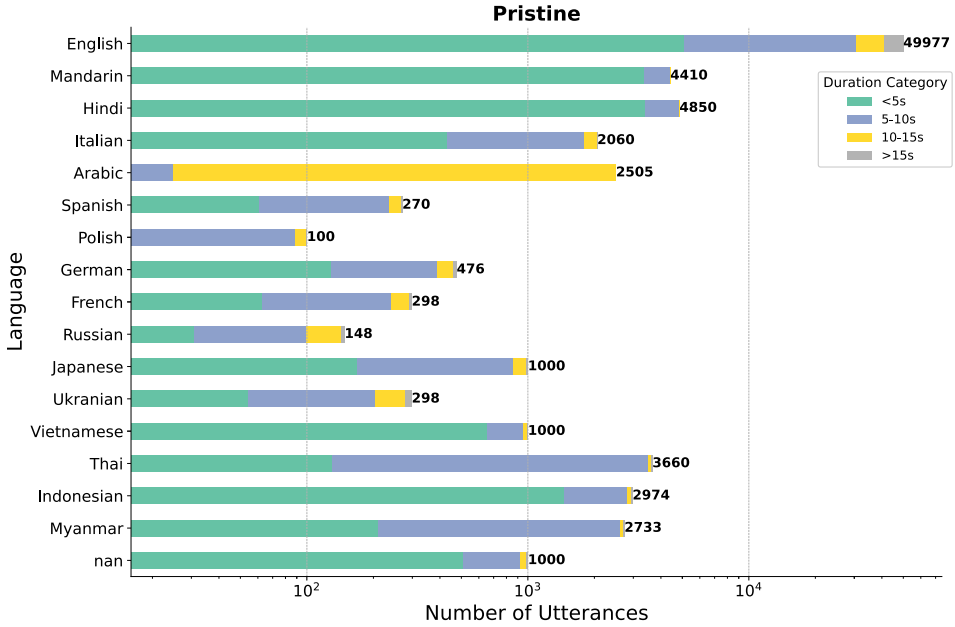

FIGURE 10. A detailed breakdown of utterance duration distribution across different languages for pristine samples in the JMDS dataset, illustrated on a logarithmic scale.

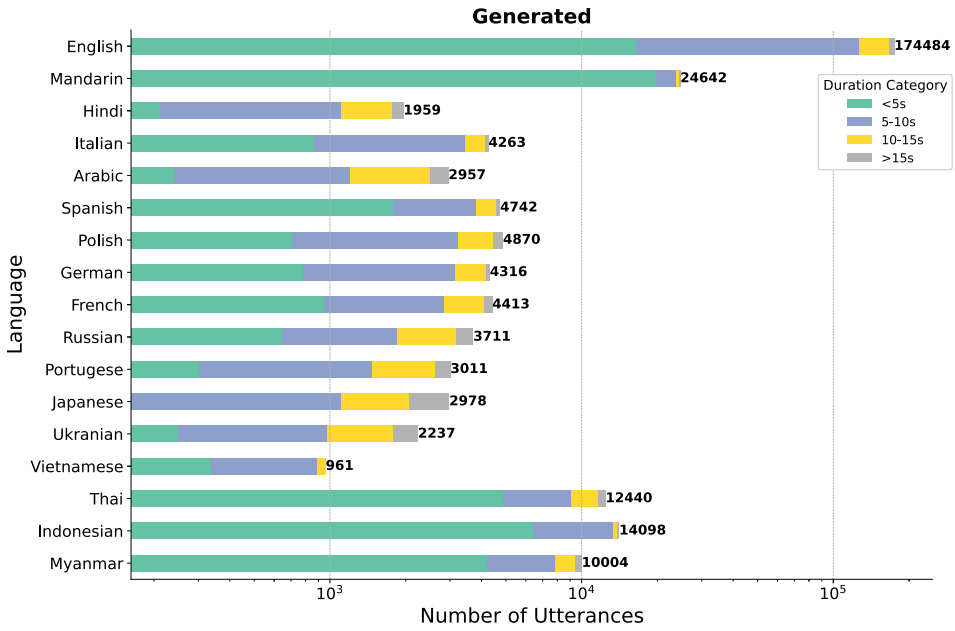

FIGURE 11. A detailed breakdown of utterance duration distribution across different languages for generated samples in the JMDS dataset, illustrated on a logarithmic scale.

2) Second, the choice of padding length in AASIST-based architectures significantly impacts detection performance—particularly on unseen data—as evidenced by our results in the SAFE challenge.
3) Third, comparative analysis of front-end and back-end architectures indicates that while Self-Supervised Learning (SSL) features enhance overall performance, AASIST-based back-ends consistently outperform ResNet and RawNet3. Notably, the superior performance of XLS-R over WavLM underscores the necessity of task-specific feature selection in anti-spoofing.
4) Finally, the language-specific analysis on JMDS-Open identified significant variations in detection accuracy across languages. In particular, Mandarin Chinese emerged as a unique challenge, likely due to the lower

acoustic quality of synthetic speech samples available in the training set.

These findings collectively underscore the complexities of developing robust, generalizable deepfake detection systems capable of navigating diverse data distributions and linguistic variations. Theoretically, AI-driven speech synthesis is language-agnostic; consequently, spoofing countermeasures should ideally transcend linguistic boundaries.

From this perspective, performance disparities across multilingual datasets should primarily arise from variations in training data conditions rather than the linguistic features themselves. However, our findings—particularly the language-specific evaluations—reveal while synthesis algorithms are theoretically language-independent, their acoustic artifacts vary significantly in practice. As evidenced by the detection challenges in Mandarin Chinese, the maturity of AI-generated speech in certain languages has not yet reached parity with English. This disparity introduces a latent language dependency, where artifact prominence is dictated by the acoustic properties and synthesis sophistication specific to each language. Consequently, while language-agnostic countermeasures remain the long-term goal, the current landscape necessitates accounting for varying synthesis maturity across different locales.

Several limitations regarding the interplay of language and synthesis quality persist. First, the language-dependent performance observed in Mandarin indicates that model effectiveness remains partially coupled with linguistic factors, likely due to varying generation quality across the dataset. Second, the uneven distribution of utterances within the JMDS dataset may hinder the development of perfectly balanced, language-agnostic models. Third, cross-dataset evaluation was restricted by inconsistent labeling across external benchmarks, precluding direct evaluation on the ASVspoof 2024 test set. Finally, the reliance on general hyperparameter configurations suggests that dataset-specific optimization could yield further gains. Future research will focus on enhancing linguistic balance within JMDS, leveraging self-supervised learning for improved generalization, and exploring refined tuning strategies.

VIII. CONCLUSION

In this work, we introduced the JMDS dataset as a valuable multilingual resource to directly address the need for diverse data to build robust deepfake speech detectors. Our evaluations demonstrated its utility for benchmarking and training models, highlighting the impact of factors such as data partitioning, input processing, and model architecture. The cross-dataset analysis, particularly the significant performance drop observed in novel linguistic and acoustic contexts, underscores the fundamental challenge of generalization to unseen deepfake attack methods. The struggle to maintain consistent performance across different languages and diverse datasets is a critical indicator of a model's reliance on dataset-specific artifacts, rather than the intrinsic, method-agnostic features of deepfake speech. Ultimately, our study contributes to a deeper understanding of the current state of detection and provides a foundation for future research aimed at developing effective countermeasures that are robust to completely unknown deepfake synthesis techniques.

DATA AVAILABILITY STATEMENT

The **J**AIST **M**ultilingual **D**eepfake **S**peech (JMDS) dataset comprises data from both publicly available repositories and private sources, with a comprehensive list of public datasets and relevant citations provided in Section III. While proprietary restrictions prevent the full dataset from being released publicly, we have made the JMDS-Open generated data and sample source code for dataset preparation (available at https://candyolivia.github.io/research/speech_security/JMDS/) accessible to the research community to support reproducibility. Aggregated statistics and detailed analyses are further included within this manuscript to support our findings, and researchers are encouraged to utilize these publicly available resources for replication efforts or contact the corresponding author for further inquiries.

ETHICAL CONSIDERATIONS

The collection of data for the JMDS dataset involved both open-source and private sources, necessitating careful ethical considerations. Open-source data was utilized from publicly available repositories, adhering to their respective licenses and terms of use. For the private data component, we prioritized ethical acquisition, ensuring that all data was collected appropriately and with explicit approval obtained prior to its inclusion in the dataset. This process aimed to respect privacy and comply with relevant data handling regulations. By implementing this dual approach, we sought to create a comprehensive and diverse dataset while upholding ethical standards in our data sourcing practices.

ACKNOWLEDGMENT

The authors extend their sincere thanks to their 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.

(Candy Olivia Mawalim, Yutong Wang, and Aulia Adila contributed equally to this work.)

APPENDIX A
DETAILS RELATED TO THE JMDS DATASET

Table 8 and 9, Figure 8–11

REFERENCES

[1] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, ''Audio deepfake detection: A survey,'' 2023, *arXiv:2308.14970*.

[2] X. Wang et al., ''ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,'' *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[3] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. V. Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, ''XLS-R: Self-supervised cross-lingual speech representation learning at scale,'' in *Proc. Interspeech*, Sep. 2022, pp. 2278–2282.

[4] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5621–5625.

[5] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *J. Mach. Learn. Res.*, vol. 25, pp. 97:1–97:52, 2024. [Online]. Available: https://jmlr.org/papers/v25/23-1318.html

[6] Z. Ba, Q. Wen, P. Cheng, Y. Wang, F. Lin, L. Lu, and Z. Liu, "Transferring audio deepfake detection capability across languages," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 2033–2044.

[7] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, 2015, pp. 2037–2041.

[8] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Proc. Speaker Lang. Recognit. Workshop*, 2018, pp. 296–303.

[9] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *Proc. SpeD*, Oct. 2019, pp. 1–10.

[10] J. H. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio DeepFake detection," in *Proc. NeurIPS Track Datasets Benchmarks*, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract-round2.html

[11] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. Aik Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," 2021, *arXiv:2109.00537*.

[12] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S.-M. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "ADD 2022: The first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022, pp. 9216–9220.

[13] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese dataset for fake audio detection," *Speech Commun.*, vol. 164, Oct. 2024, Art. no. 103122.

[14] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech*, 2022, pp. 2783–2787.

[15] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, L. Shan, Z. Lian, S. Nie, and H. Li, "ADD 2023: The second audio deepfake detection challenge," in *Proc. Workshop Deepfake Audio Detection Anal. Co-located*, Aug. 2023, pp. 125–130. [Online]. Available: https://ceur-ws.org/Vol-3597/paper21.pdf

[16] J. J. Bird and A. Lotfi, "Real-time detection of AI-generated speech for DeepFake voice conversion," 2023, *arXiv:2308.12734*.

[17] N. M. Müller, P. Kawa, W. Herng Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "MLAAD: The multi-language audio anti-spoofing dataset," 2024, *arXiv:2401.09512*.

[18] Y. Li, M. Zhang, M. Ren, X. Qiao, M. Ma, D. Wei, and H. Yang, "Cross-domain audio deepfake detection: Dataset and analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Miami, FL, USA, Nov. 2024, pp. 4977–4983. [Online]. Available: https://aclanthology.org/2024.emnlp-main.286

[19] X. Wang et al., "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," 2025, *arXiv:2502.08857*.

[20] J. Du, I.-M. Lin, I.-H. Chiu, X. Chen, H. Wu, W. Ren, Y. Tsao, H.-Y. Lee, and J.-S. R. Jang, "DFADD: The diffusion and flow-matching based audio deepfake dataset," in *Proc. SLT*, 2024, pp. 921–928.

[21] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "SafeEar: Content privacy-preserving audio deepfake detection," in *Proc. ACM SIGSAC*, 2024, pp. 3585–3599, doi: 10.1145/3658644.3670285.

[22] T. Kirill, P. Cummer, P. Pherwani, J. Aslam, M. Davinroy, P. Bautista, L. Cassani, and M. C. Stamm, "SAFE: Synthetic audio forensics evaluation challenge," in *Proc. ACM IH MMSEC*, 2025, pp. 174–180.

[23] K. Ito and L. Johnson. (2017). *The Lj Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[24] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017, *arXiv:1711.00354*.

[25] İmdat Celeste. (2020). *The M-AILABS Speech Dataset*. Accessed: Apr. 2025. [Online]. Available: https://github.com/imdatceleste/m-ailabs-dataset

[26] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[27] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023.

[28] X. Wang, H. Delgado, H. Tak, J. Jung, H. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. W. D. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," 2024, *arXiv:2408.08739*.

[29] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, Oct. 2020, pp. 2757–2761.

[30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, Nov. 2017, pp. 1–5.

[31] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus," in *Proc. Interspeech*, Aug. 2021, pp. 2756–2760.

[32] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proc. Interspeech*, 2021, pp. 3665–3669.

[33] C. O. Mawalim, K. Galajit, D. P. Lestari, W. P. Pa, and M. Unoki, "Challenges in speech spoofing countermeasures for Southeast Asian languages," Acoust. Soc. Japan (ASJ), Tokyo, Japan, Tech. Rep. 3-P-4, 2025.

[34] K. Galajit, T. Kosolsriwiwat, M. Unoki, C. O. Mawalim, P. Aimmanee, W. Kongprawechnon, W. P. Pa, A. Chaiwongyen, T. Racharak, S. Boonkla, H. Yassin, and J. Karnjana, "ThaiSpoof: A database for spoof detection in Thai language," in *Proc. (iSAI-NLP)*, Nov. 2023, pp. 1–6.

[35] S. A. Arief, C. O. Mawalim, and D. P. Lestari, "Indonesian speech anti-spoofing system: Data creation and convolutional neural network models," in *Proc. 11th Int. Conf. Adv. Inform., Concept, Theory Appl. (ICAICTA)*, Sep. 2024, pp. 1–6.

[36] C. O. Mawalim, S. A. Arief, and D. P. Lestari, "InaSAS: Benchmarking Indonesian speech antispoofing systems," *APSIPA Trans. Signal Inf. Process.*, vol. 14, no. 3, pp. 1–44, Jun. 2025, doi: 10.1561/116.20240080.

[37] V. Hoang, V. T. Pham, H. N. Xuan, P. Nhi, P. Dat, and T. T. T. Nguyen, "VSASV: A Vietnamese dataset for spoofing-aware speaker verification," in *Proc. Interspeech*, Sep. 2024, pp. 4288–4292.

[38] H. M. S. Naing, W. P. Pa, A. M. Hlaing, M. A. A. Aung, K. Galajit, and C. O. Mawalim, "UCSYSpoof: A Myanmar language dataset for voice spoofing detection," in *Proc. 27th Conf. Oriental COCOSDA Int. Committee Co-Ordination Standardisation Speech Databases Assessment Techn. (O-COCOSDA)*, Oct. 2024, pp. 1–5.

[39] R. Kolobov, O. Okhapkina, O. Omelchishina, A. Platunov, R. Bedyakin, V. Moshkin, D. Menshikov, and N. Mikhaylovskiy, "MediaSpeech: Multilanguage ASR benchmark and dataset," 2021, *arXiv:2103.16193*.

[40] T. Javed et al., "IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages," in *Proc. Findings ACL*, Aug. 2024, pp. 10740–10782.

[41] A. C. Junior, E. Casanova, A. Soares, F. S. de Oliveira, L. Oliveira, R. C. F. Junior, D. P. P. da Silva, F. G. Fayet, B. B. Carlotto, L. R. S. Gris, and S. M. Aluísio, "CORAA ASR: A large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese," *Lang. Resour. Eval.*, vol. 57, no. 3, pp. 1–33, Sep. 2023.

[42] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," 2019, *arXiv:1908.06248*.

[43] National Electronics and Computer Technology Center (NECTEC). *Lotus (Thai Speech Recognition Corpus)*. Accessed: Nov. 20, 2025. [Online]. Available: https://nlpforthai.com/tasks/speech-recognition/

[44] H.-T. Luong and H.-Q. Vu, "A non-expert kaldi recipe for Vietnamese speech recognition system," in *Proc. 3rd Int. Workshop Worldwide Lang. Service Infrastruct. 2nd Workshop Open Infrastructures Anal. Frameworks Human Lang. Technol. (WLSI/OIAF4HLT)*, Dec. 2016, pp. 51–55. [Online]. Available: https://aclanthology.org/W16-5207/

[45] A. Adila, C. O. Mawalim, and M. Unoki, "Detecting spoof voices in Asian non-native speech: An Indonesian and Thai case study," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2024, pp. 1–6.

[46] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6493–6497.

[47] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. Interspeech*, Sep. 2019, pp. 1078–1082.

[48] J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 6367–6371.

[49] J. C. Brown, "Calculation of a constant q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[50] J.-W. Jung, Y. Kim, H.-S. Heo, B. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022, pp. 2228–2232.

[51] H. Delgado, N. Evans, J.-w. Jung, T. Kinnunen, I. Kukanov, K. A. Lee, X. Liu, H.-J. Shim, M. Sahidullah, H. Tak, M. Todisco, X. Wang, and J. Yamagishi. (2024). *ASVspoof 5 Evaluation Plan*. [Online]. Available: http://www.asvspoof.org/

[52] C. O. Mawalim, Y. Wang, A. Adila, S. Okada, and M. Unoki, "Robust multilingual audio deepfake detection through hybrid modeling," in *Proc. ACM IH MMSEC*, 2025, pp. 181–192, doi: 10.1145/3733102.3736706.

[53] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

CANDY OLIVIA MAWALIM (Member, IEEE) received the B.S. degree in computer science from Institut Teknologi Bandung (ITB), Indonesia, and the M.S. and Ph.D. degrees from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), with the Ph.D. being awarded, in 2022. She was a Japan Society for the Promotion of Science (JSPS) Research Fellow for Young Scientists (DC1), from 2020 to 2022. Since April 2022, she has been a Faculty Member with the School of Information Science, JAIST, where she is currently a Senior Lecturer. Her main research interests include speech signal processing, hearing perception, voice privacy protection, and machine learning. She also serves on the Education Team for the ISCA Special Interest Group of Security and Privacy in Speech Communication (SIG-SPSC) Committee.

YUTONG WANG received the B.S. degree in computer science and technology from Dalian University of Foreign Languages, Dalian, China, and the M.S. degree from Japan Advanced Institute of Science and Technology (JAIST). His primary research interests include multi-modal deep learning, machine learning, and artificial intelligence, with a focus on integrating various data modalities for sports analytics and related applications.

AULIA ADILA received the B.S. degree in computer science from Institut Teknologi Bandung (ITB), Indonesia, and the M.S. degree in speech processing from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Japan. Her research interests include auditory signal processing, speech privacy, speech synthesis, and secure speech technologies, such as speech watermarking and spoof/deepfake detection, leveraging deep learning methods.

SHOGO OKADA (Member, IEEE) received the Ph.D. degree from Tokyo Institute of Technology, Japan, in 2008. He directs the Social Signal and Interaction Group, Japan Advanced Institute of Science and Technology (JAIST), Japan, and is a Professor with JAIST. He joined Kyoto University, as a Project Assistant Professor, in 2008, Tokyo Institute of Technology, as a Tenured Assistant Professor, in 2011, and the IDIAP Research Institute, Switzerland, as a Visiting Faculty Member, in 2014. His research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of ACM.

MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a Visiting Researcher, from 1999 to 2000, and was then a Visiting Research Associate with the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been a Faculty Member with the School of Information Science, JAIST, since 2001, where he is currently a Professor. His research interests include auditory-motivated signal processing and the modeling of auditory systems. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information, and Communication Engineers (IEICE), Japan, and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). He received the Sato Prize for an Outstanding Paper from ASJ, in 1999, 2010, and 2013, and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.