Multilingual Deepfake Speech Dataset for Robust and Generalizable Detection

Candy Olivia Mawalim*, *Member, IEEE,* Yutong Wang*, Aulia Adila*, Shogo Okada, *Member, IEEE,* Masashi Unoki, *Member, IEEE*

***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 400,000 utterances, JMDS incorporates a wide range of deepfake generation methods and includes both human (pristine) and machine-generated 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, multilingual dataset, generalizability**

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], [3].

An emerging defense strategy against deepfake speech is spoofing detection, which aims to distinguish between human 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 [4]–[6], it becomes increasingly critical to advance spoofing detection capabilities beyond high-resource languages to ensure inclusive protection. While 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 [7].

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 multisource 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 compre-

[*]These authors contributed equally.

TABLE I: 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**	**No. of Utts.**	**No. of Pristine**	**No. of Generated**	**Focus task(s)**
ASVspoof2015 [8]	2015	English	TTS, Vocoder, VC	263,151	16,651	246,500	Spoofing detection (speech synthesis and voice conversion)
ASVspoof2017 [9]	2017	English	Replay	18,030	3,565	14,465	Spoofing detection (replay attacks)
Fake or Real [10]	2019	English	TTS	>198,029	>110,744	87,285	General deepfake speech detection
ASVspoof2019 [3]	2019	English	TTS, Vocoder, VC, Replay	339,891	41,373	298,518	Spoofing detection (speech synthesis, voice conversion, and replay attacks)
WaveFake [11]	2021	English, Japanese	TTS, Vocoder	117,985	0*	117,985	General deepfake speech detection
ASVspoof2021 [12]	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 [13]	2022	Mandarin	TTS, VC, Partially Fake	53,577	5,619	47,958	General deepfake speech detection
CFAD [14]	2022	Mandarin	TTS, Partially Fake	347,400	115,800	231,600	General deepfake speech detection (robustness, generalization)
In-the-Wild [15]	2022	English	TTS	31,779	19,963	11,816	Real-world deepfake speech detection
ADD2023 [16]	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 [17]	2023	English	VC (retrieval-based)	7,484	3,742	3,742	Voice conversion deepfake detection
DECRO [7]	2023	English, Mandarin	TTS, VC	118,381	33,702	84,679	Cross-language deepfake speech detection
MLAAD [18]	2024	Multilingual (38 lang.)	TTS, Vocoder	154,000	0*	154,000	Multilingual deepfake speech detection
CD-ADD [19]	2024	English	TTS	145,570	25,111	120,459	Cross-domain deepfake speech detection
ASVspoof5 [20]	2024	English	TTS, Vocoder, VC, AT	>1,293,892	252,050	>1,041,842	General deepfake speech and spoofing detection (including adversarial attacks)
DFADD [21]	2024	English	TTS	207,955	44,455	163,500	Spoofing detection (diffusion and flow-matching based TTS)
CVoiceFake [22]	2024	Multilingual (5 lang.)	Vocoder	1,254,893	0*	1,254,893	Multilingual deepfake speech detection and speech content privacy preservation
SAFE Challenge [23]	2025	Multilingual	Unknown	Unknown	Unknown	Unknown	Unseen and various deepfake speech detection (robustness)
JMDS-Open (ours)	2025	Multilingual (15 lang.)	TTS, Vocoder, VC, AT	312,146	65,546	246,600	Multilingual deepfake speech detection
JMDS-All (ours)	2025	Multilingual (17 lang.)	TTS, Vocoder, VC, AT	412,021	93,968	318,053	Multilingual deepfake speech detection

* Pristine data are sourced from LJSpeech [24] and JSUT [25] (WaveFake); M-AILABS [26] (MLAAD); and CommonVoice [27] (CVoiceFake).

hensive 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 [3], [8], [9], [28], [29], 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 [29], is built on the MLS [30] 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 [13], [16], which addresses more complex real-world detection scenarios. The first edition focused on detecting low-quality and partially fake audio [13], while the second edition expanded to include manipulation localization and generation method identification [16]. The ADD dataset is derived from Mandarin corpora, AISHELL-1 [31], AISHELL-3 [32], and AISHELL-4 [33], and contains samples generated using a range of TTS and VC models, though the specific models are not publicly disclosed.

Filling the gap, CFAD [14] 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 [10], introduced in 2019, includes samples generated using a combination of open-source and commercial tools. The In-the-Wild dataset [15], 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 [21], which contains deepfake speech generated using five state-of-the-art diffusion and flow-matching TTS models. To

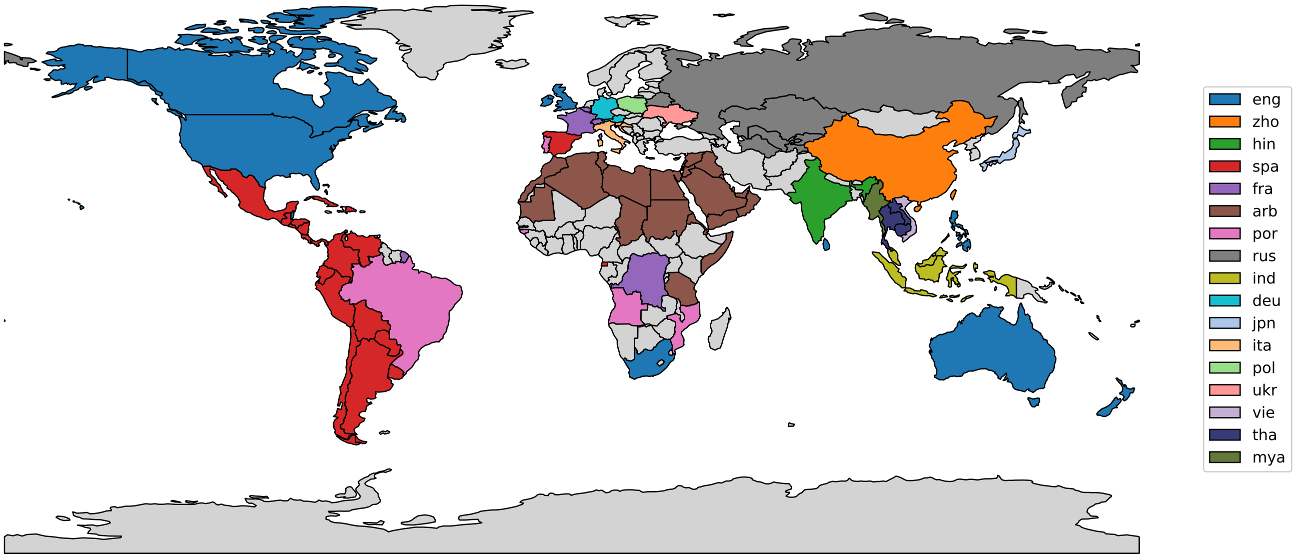

Fig. 1: World map illustrating the primary geographical regions where the 17 languages included in our multilingual corpus are spoken as *de facto* and/or *de jure* languages. The legend provides the corresponding ISO 639 codes for each region.

address the growing challenges posed by zero-shot TTS systems, the Cross-Domain Audio Deepfake Detection (CD-ADD) dataset [19] was developed to support detection models across varying domains. A separate dataset utilizing retrieval-based voice conversion systems such as DEEP-VOICE [17] was also introduced. Furthermore, to support cross-lingual evaluation of detection systems, the DECRO dataset [7] 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 [11] includes generated speech synthesized using neural vocoder models, covering both English and Japanese. A prominent multilingual dataset is MLAAD [18], which spans 38 languages and is built on the M-AILABS dataset [26], 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 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 [22], which contains over 1.25 million bonafide and deepfake utterances across five languages, with speech data sourced from the CommonVoice dataset [27].

Recent efforts have also focused on detecting deepfake speech in low-resource languages. Notably, studies have targeted languages within the ASEAN region [34], aiming to develop spoofing countermeasures tailored to these linguistic contexts. This line of work includes the construction of dedicated datasets for Thai (ThaiSpoof [35]), Indonesian (InaSpoof [36], [37]), Vietnamese (VSASV [38]), and Burmese (UCSYSpoof [39]). 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 I 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. Moreover, few incorporate both pristine and generated speech from a wide variety of sources, which is critical for building generalizable and robust detection models.

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, as shown in Fig. 2, 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.

TABLE II: Overview of data sources used in the JMDS dataset, detailing the adopted languages, data types (pristine and/or generated), deepfake methods (TTS/text-to-speech, vocoder, VC/voice conversion, AT/adversarial attack), recording or synthesis conditions, audio format, and sample rate. “Multiple” in format and sample rate means that the original speech recording of the source data contains audio with more than one format or sample rate.

Source	Adopted language(s)	Type(s)	Deepfake methods	Audio condition	Format	Sample rate
ASVspoof [20]	English	Pristine, Generated	TTS, Vocoder, VC, AT	Clean with post processing	FLAC	16 kHz
ADD [13]	Mandarin	Generated	TTS, VC	Noisy	WAV	16 kHz
MLAAD [18]	Hindi, Spanish, French, Arabic, Thai Portuguese, Russian, German, Japanese, Italian, Polish, Ukrainian, Vietnamese	Generated	TTS, Vocoder	Clean	WAV	22 kHz
AISHELL-3 [32]	Mandarin	Pristine	-	Clean	WAV	16 kHz
M-AILABS [26]	Spanish, French, Russian, German, Italian, Polish, Ukrainian	Pristine	-	Unknown	WAV	16 kHz
MediaSpeech [40]	Arabic	Pristine	-	Unknown	WAV	16 kHz
Indic-voice [41]	Hindi	Pristine	-	Clean, Noisy	WAV	8 kHz
Private data	Burmese, Indonesian, Vietnamese English (nonnative), Japanese, Thai	Pristine, Generated	TTS, Vocoder, VC	Clean, Noisy	Multiple	Multiple

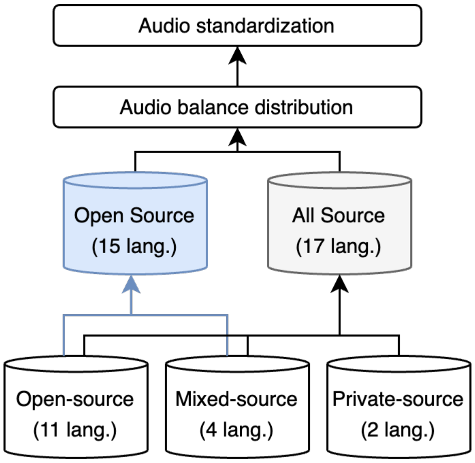

Fig. 2: JMDS dataset curation process, comprising the source corpora selection and acquisition, balance distribution, and standardization.

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 II 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 noise levels, audio formats, and sampling rates.

For English, we utilized well-established corpora from the ASVspoof Challenge [29] 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 [30] 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 [13] to supplement generated data. The pristine Chinese data was sourced from AISHELL-3 [32], 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 incorporated established corpora such as M-AILABS [26] and MLAAD [18]. In addition, language-specific resources like IndicVoice [41] and MediaSpeech [40] were used to supplement pristine data in the corresponding languages, i.e. Hindi and Arabic.

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 and also apply to languages such as Japanese, Vietnamese, and Thai. The inclusion of these recordings helped ensure broad linguistic coverage and representativeness.

Private-source data were also incorporated from prior studies on spoofing detection in Asian languages [34]–[37], [39], [42], 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

TABLE III: Summary of the number of utterances (no. of utts.) distribution across 17 languages in the JMDS dataset.

Language	**Open Source (no. of utts.)**		**All Source (utts.)**	
	Pristine	**Generated**	**Pristine**	**Generated**
English (eng)	50,131	175,040	51,931	175,040
Mandarin (zho)	4,410	30,267	4,410	30,267
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)	0	3,011	0	3,011
Japanese (jpn)	0	2,978	5,037	2,978
Ukrainian (ukr)	298	2,237	298	2,237
Vietnamese (vie)	0	961	4,968	961
Thai (tha)	0	875	5,263	24,305
Indonesian (ind)	0	0	5,863	27,960
Burmese (mya)	0	0	5,491	20,063
Total	**65,546**	**246,600**	**93,968**	**318,053**
Subtotal	**312,146**		**412,021**	

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 three 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 III.

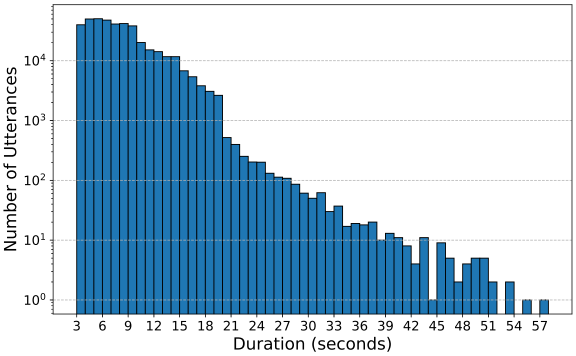

Fig. 3: Histogram showing the distribution of utterance durations in the JMDS dataset (All Source) on a logarithmic scale.

During the final standardization phase, we limited each audio clip to a maximum of 60 seconds to ensure manageability and consistency during training and evaluation. All audio files were resampled to 16 kHz and converted to monochannel format. We also verified audio integrity by checking for file corruption and empty clips, and ensuring compliance with the .wav format. Conducting such a meticulous data curation process has allowed us to compile a well-structured, multilingual speech corpus that is consistent, diverse, and suitable for advancing research in deepfake speech detection.

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 412,021 speech samples across 17 languages, including English and Chinese, which represent some of the world’s most widely spoken languages.

To analyze the temporal characteristics of the dataset, we examined the distribution of utterance durations, as depicted in Fig. 3. Most samples fall within the range of approximately 3 to 10 seconds. A breakdown of utterance duration distributions across different languages for both pristine and generated data is shown in Figs. 9 and 10. 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 [20].

We incorporated generated speech samples from four different data sources: ASVspoof [29], ADD [13], MLAAD [18], 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 contains generated samples produced utilizing commonly used TTS and VC algorithms, although the exact models are not disclosed. Our private-source generated speech was created using high-performance speech synthesis systems, encompassing TTS, vocoders, VC, and several proprietary TTS models.

A comprehensive list of the generation algorithms included in the JMDS dataset, along with their corresponding data sources, is presented in Table IX in the Appendix. 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). Additionally, for clarity, we grouped certain distinct algorithms that are built on the same base model under a single category.

D. Audio Quality

We analyze the quality of the audio data using the Mean Opinion Score (MOS) obtained from DNSMOS [43], 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 subjective human evaluation, which is considered the "gold standard" in assessing speech quality optimized for human perception. 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. A detailed breakdown of the average MOS scores per language is provided in Table VIII in the Appendix, while a comparative view of MOS score distributions across data sources is shown in Figs. 7 and 8 in the Appendix.

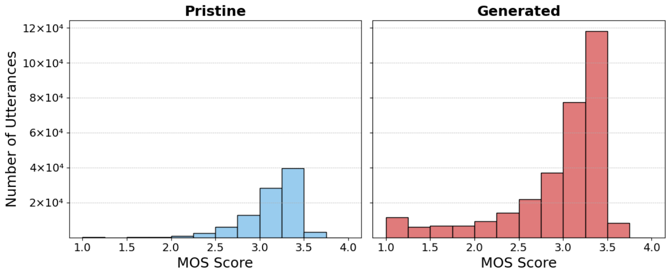

Fig. 4: Distribution of MOS scores in the JMDS dataset (All Source).

Figure 4 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 [44] and AASIST-based [45] 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 subtle, 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 [44] and used Constant-Q Transform (CQT) features [46] 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 [45]. 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 subtle and informative acoustic patterns.

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 [47]. 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 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.

B. Data Partition

We validate our dataset design and collection through experiments in building a robust and generalized spoofing detection.

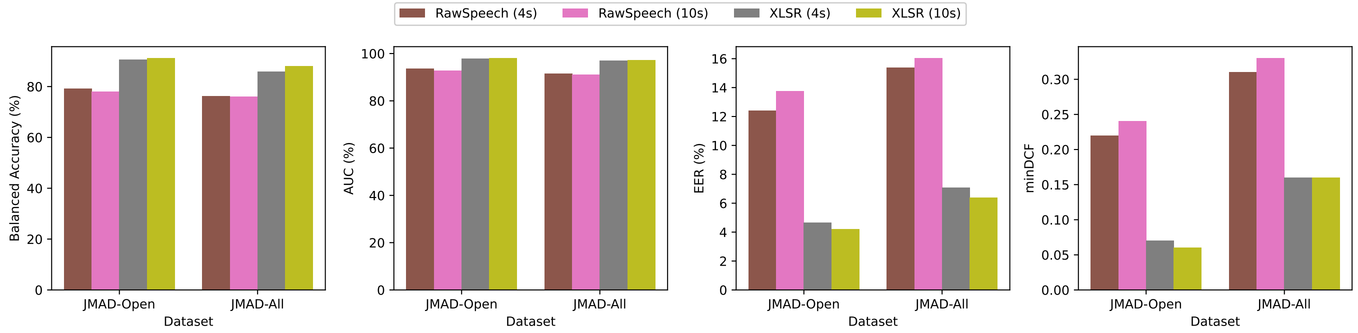

Fig. 5: Performance comparison of AASIST-based models with 4-second and 10-second padding.

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

Partition	**Subset**	**No. of 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

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 when this information is available. This process ensures that each subset maintains a similar distribution of these attributes. Furthermore, the three subsets are mutually exclusive, with no overlapping speech samples between them.

Table IV 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 that caused the total count in the "All" partition to differ from Table III. These problematic utterances 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.

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 V. All method pairs in this evaluation were trained utilizing a default audio padding of 4 seconds.

In our initial cross-partition evaluation on the JMDS dataset, we observed a significant performance disparity across different model architectures. The classical approach utilizing CQT features with a ResNet34 backbone struggled to accurately detect pristine signals, often classifying all speech signals as generated, resulting in low balanced accuracy and AUC, and a high EER. In contrast, the end-to-end AASIST architecture demonstrated superior performance. Utilizing raw waveforms (RawSpeech), it achieved a balanced accuracy of approximately 79.15%, an AUC of 93.51%, and an EER of 12.39% when both training and evaluation were performed on JMDS-Open. Notably, the integration of SSL features, particularly with XLS-R (300M), further enhanced the results, yielding a balanced accuracy of 90.66%, an AUC of 97.86%, and an EER of 4.67%. The performance trends remained consistent when comparing different front-end and back-end model pairings.

Furthermore, our analysis of the Open and All partitions revealed that training on JMDS-Open generally led to better performance compared to training on JMDS-All. This can potentially be attributed to the inclusion of controlled-environment recordings and a low-resource language within the private data of the "All" partition, which increases the difficulty of the detection task. Consequently, the JMDS-All dataset appears to be a more rigorous benchmark for evaluating the robustness of deepfake speech detection models.

In our subsequent analysis, we investigated the impact of different padding lengths on models built using the AASIST architecture. Our experiments revealed that the chosen padding duration is often crucial for accurately detecting artifacts in generated speech, with its importance depending on the characteristics of the evaluation set. For example, a shorter padding might suffice if the evaluation data primarily consists of short utterances (under 5 seconds). Given the diverse utterance lengths within our JMDS dataset, we compared models trained with 4-second and 10-second padding. Figure 5 illustrates the performance of these models, evaluated in terms of balanced accuracy, AUC, EER, and minDCF, on both JMDS-Open and JMDS-All, with training performed on JMDS-Open. Notably, when using raw waveforms as input, the difference in performance between the two padding lengths was minimal. However, we observed a slight improvement when utilizing XLS-R features with the longer 10-second padding. To further evaluate the generalization capabilities

TABLE V: Cross-partition evaluation on the JMDS dataset, comparing models trained on the OpenSource subset versus the All Source 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	ResNet34	JMDS-Open	Pristine	0.00	39.99	23.21	69.69	0.00	39.24	24.64	68.10
			Generated	79.98				78.47			
RawSpeech	AASIST		Pristine	59.32	79.15	93.51	12.39	54.27	76.23	91.49	15.39
			Generated	98.97				98.19			
XLS-R (300M)	AASIST		Pristine	**81.52**	**90.66**	**97.86**	**4.67**	**74.88**	**85.95**	**97.02**	**7.06**
			Generated	**99.79**				**97.02**			
CQT	ResNet34	JMDS-All	Pristine	0.00	39.99	24.62	68.49	0.00	39.24	20.60	71.24
			Generated	79.98				78.47			
RawSpeech	AASIST		Pristine	60.17	79.78	93.94	11.96	66.02	82.56	94.83	10.44
			Generated	99.40				99.09			
XLS-R (300M)	AASIST		Pristine	79.76	89.77	97.59	5.02	73.51	85.83	96.41	7.98
			Generated	99.78				98.15			

TABLE VI: 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	25.03	54.19	58.58	44.09	1.00
		Generated	83.34				
XLS-R (300M)	ResNet	Pristine	43.76	68.88	84.96	22.65	0.58
		Generated	94.01				
RawSpeech	AASIST	Pristine	57.17	78.01	92.73	13.75	0.24
		Generated	98.85				
WavLM	AASIST	Pristine	62.12	80.09	95.22	10.99	0.26
		Generated	98.06				
XLS-R (300M)	AASIST	Pristine	82.39	91.14	98.01	4.21	0.06
		Generated	99.88				

TABLE VII: 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 [10]		ASVspoof2019 [3]		ITW [15]		DECRO [7]		JMDS-Open		SAFE [23]
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	54.46	23.85	74.75	44.97	53.74	48.58
RawSpeech	AASIST		12.51	88.13	**2.42**	**93.73**	39.59	62.51	35.18	71.81	35.96	59.25	42.70
CQT	ResNet	ASVspoof2024	33.39	65.36	23.82	60.98	36.34	65.09	34.50	61.20	44.92	56.76	36.73
RawSpeech	AASIST		7.15	92.39	21.76	63.20	16.39	83.76	35.77	79.57	23.15	68.71	47.85
CQT	ResNet	JMDS-Open	18.44	81.94	23.52	59.78	45.72	53.12	29.15	68.11	44.09	54.19	58.74
RawSpeech	AASIST		**0.48**	**99.57**	19.77	62.94	7.01	93.24	24.90	75.24	13.75	78.01	**67.27**
XLS-R	AASIST		2.49	97.97	10.03	78.17	**1.92**	**98.08**	**13.26**	**86.35**	**4.21**	**91.14**	60.73

of our models on entirely unseen data, we assessed their performance on the SAFE challenge 2025[1] [23] [48]. Utilizing the methods compared in this study, we observed a significant performance improvement (over 5% in balanced accuracy) with increased padding size when tested on this completely unknown dataset, which included audio samples with varying lengths up to 60 seconds. A more detailed analysis of our cross-dataset evaluation is presented in Section VI.

To provide a fair comparison of performance across different front-end and back-end method pairings, we trained models on the JMDS-Open dataset using a 10-second padding and evaluated them on the same dataset. The results of this analysis are presented in Table 7, showcasing five combinations: (1) CQT-ResNet, (2) XLSR-ResNet, (3) RawSpeech-AASIST, (4) WavLM-AASIST, and (5) XLSR-AASIST. All evaluation metrics detailed in Subsection V-A were utilized. Initially, when comparing the traditional CQT features against features extracted from the pre-trained SSL model XLS-R (300M, abbreviated as XLSR), both with a ResNet back-end, we observed a substantial performance difference. Utilizing the pre-trained SSL features yielded more robust results, albeit at a higher computational cost. Subsequently, the AASIST-based models generally outperformed their ResNet counterparts. For example, with XLSR as the front-end, AASIST achieved a balanced accuracy of approximately 91.14%, significantly higher than the 78.01% obtained by the ResNet model. Furthermore, our findings highlight the importance of selecting the appropriate SSL model. While WavLM often surpasses XLS-R in various benchmark speech processing tasks [49], in this specific deepfake speech detection scenario, XLSR demonstrated superior performance across all metrics,

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

exhibiting over 10% higher balanced accuracy, a 5% lower EER, and a 0.2 lower minDCF compared to WavLM.

VI. Cross-dataset Evaluation

To assess the generalization capabilities of our models beyond the JMDS dataset, we conducted a comprehensive cross-dataset evaluation. This involved training our models on datasets, specifically, ASVspoof 2019 [3], ASVspoof 2024 [47], and JMDS-Open. Subsequently, we evaluate their performance on these same datasets as well as the completely unseen datasets, particularly FoR [10], in-the-wild (ITW) [15], DECRO [7], and SAFE challenge 2025 [23].

Table VII presents the results of our cross-dataset evaluation. It is important to note that we could not directly evaluate our models on the ASVspoof 2024 evaluation data due to a lack of ground-truth labels. Instead, we trained models on the ASVspoof 2024 training data and assessed their performance on other datasets. For the SAFE benchmark, we were only able to report the balanced accuracy as it is the only metric available on the challenge's leaderboard.

The results shown here were obtained using general model architectures and standard hyperparameter tuning on validation sets, without any dataset-specific optimizations. The significant variability in balanced accuracy highlights the ongoing challenge of generalization in this field, especially when comparing the JMDS-Open dataset to existing benchmarks. While models often perform well when training and evaluation sets are aligned, their effectiveness significantly diminishes on unseen datasets.

Consistent with previous research, AASIST-based models generally outperformed ResNet. The results also indicate that in challenging, real-world environments with completely unknown data, detection performance still needs improvement, with accuracy on the SAFE benchmark being around 67%.

The biggest issue for improving robustness and generalization in deepfake speech detection is the challenge of evaluation across unknown attacks. While cross-lingual evaluation presents a significant obstacle, models typically exhibit a more catastrophic performance drop when encountering an unseen high-quality deepfake generation method—known as a zero-shot attack—compared to simply switching to an unseen language generated by a known method. This effect is clearly supported by the results in Table VII. The SAFE challenge utilizes highly advanced unknown attacks [23], leading to significantly lower detection accuracy. Conversely, the English-language FoR and ITW datasets, while containing unknown attacks, are significantly easier to detect (accuracy typically $>$ 90%) because their synthesis methods are less advanced (predating 2022). This contrast indicates that the sophistication of the generation method—rather than language difference alone—is the primary factor limiting detection accuracy when a reliable training benchmark is available.

The JMDS-Open dataset also proves to be 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 78.19% on ASVspoof 2019 and 67.27% on the SAFE challenge. In contrast, models trained on either of the ASVspoof datasets performed significantly worse on the SAFE challenge, with balanced accuracies falling below 50%, despite showing slightly better performance on their respective evaluation sets. This demonstrates the superior cross-dataset generalizability of models trained on JMDS-Open.

To further investigate the language-specific performance of our best-performing architecture, the RawSpeech-AASIST, we conducted an analysis on the three most prevalent languages in the JMDS-Open dataset: English (eng), Mandarin Chinese (zho), and Hindi (hin). The models used for this evaluation were trained on a combined corpus of ASVspoof 2019, ASVspoof 2024, and the entire JMDS-Open dataset. The detailed results are presented in Fig. 6.

The language-specific evaluation uncovered notable patterns in the model's generalization capabilities. As anticipated, performance was generally strongest on the model trained on JMDS-Open. The detection accuracy on Mandarin Chinese (zho) across all models, as depicted in Fig. 6, was particularly poor. While the countermeasure scores for pristine Mandarin speech evaluated using a model trained on ASVspoof 2024 and JMDS-Open leaned positive, the distribution of generated Mandarin speech significantly overlapped with that of pristine speech. This subpar detection in Mandarin Chinese could be attributed to the lower quality of generated speech in the Audio Deepfake Database (ADD) benchmark, where the Mean Opinion Score (MOS) is typically below 3.0 (as illustrated in Figs. 7 and 8). This language-specific analysis highlights areas for future improvement in multilingual deepfake speech detection.

Moving forward, our work will encompass more extensive cross-dataset evaluations across a broader spectrum of benchmarks. Furthermore, we anticipate that the insights gained from this analysis will aid in the strategic selection of training and evaluation datasets to foster better generalization in deepfake speech detection models.

VII. Discussion

This study presented the JMDS dataset, a multilingual resource for generated speech detection, and evaluated its utility through benchmark comparisons and cross-dataset experiments. Our findings highlight several key aspects.

1) First, cross-evaluation within JMDS revealed that the partition containing private data, while potentially more challenging, serves as a rigorous testbed.
2) Second, the choice of padding length in AASIST-based models significantly impacts performance, particularly on unseen data, as demonstrated by the SAFE challenge results.
3) Third, our comparison of front-end and back-end architectures indicated that while SSL-based features enhance performance, AASIST models generally outperform ResNet. Notably, the effectiveness of specific SSL models, such as the superior performance of XLS-R over WavLM in our experiments, underscores the importance of task-specific feature selection.
4) Finally, the language-specific analysis on JMDS-Open revealed variations in detection accuracy across languages,

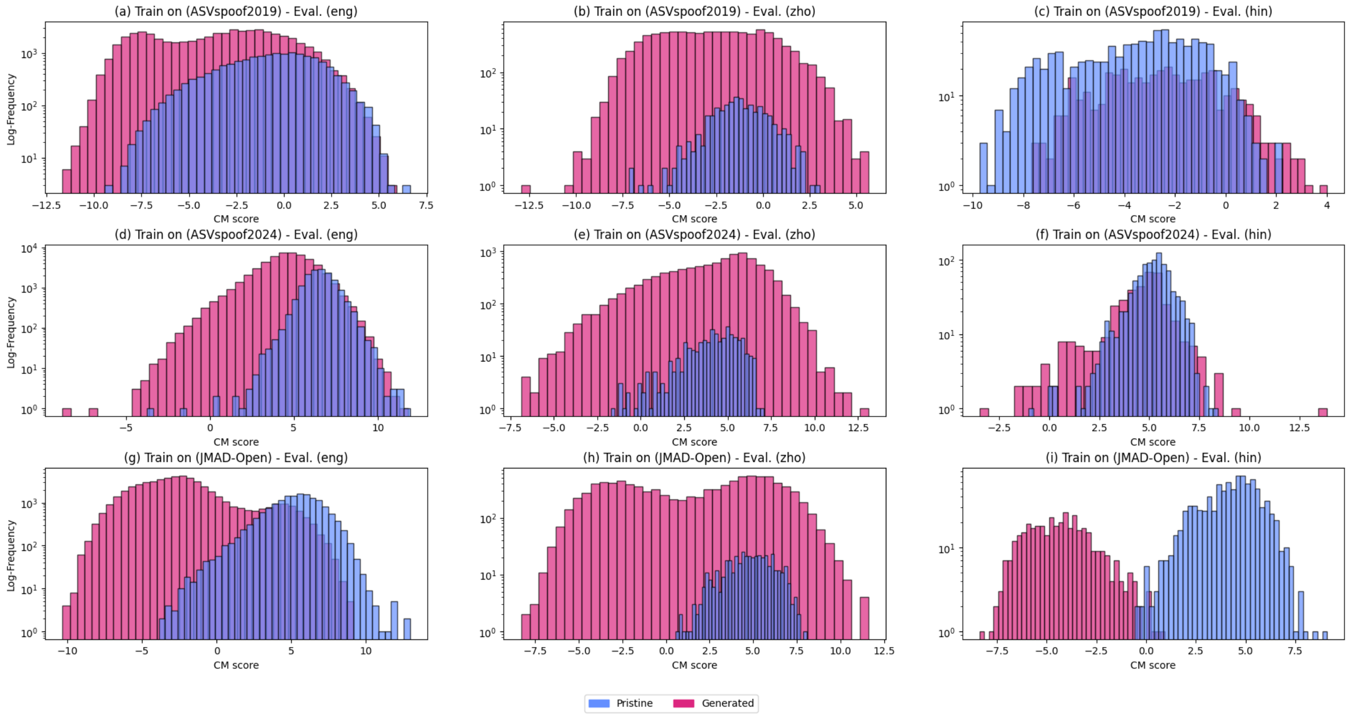

Fig. 6: Language-specific performance using RawSpeech-AASIST method. Three most represented languages within JMDS-Open were included: English (eng), Mandarin Chinese (zho), and Hindi (hin).

with Mandarin Chinese presenting a particular challenge—potentially due to the lower quality of available generated speech in the training data.

These findings collectively emphasize the complexities of building robust and generalizable deepfake speech detection systems capable of handling diverse data distributions and linguistic variations. Delving deeper into this aspect of linguistic variation, AI-based speech synthesis techniques are ideally language-agnostic, meaning that deepfake and spoofing countermeasures should similarly transcend linguistic boundaries.

From this point of view, performance variations across multilingual datasets would primarily stem from differences in training regimes and data conditions rather than the languages themselves. However, our findings, particularly the language-specific evaluation, indicate a more complex reality. While the underlying AI synthesis may be language-independent in principle, the practical quality and characteristics of synthesized speech demonstrably vary across languages. As highlighted by the challenges in detecting deepfakes in Mandarin Chinese, the current quality of AI-generated speech in certain languages is not on par with that of English. This disparity in synthesis quality introduces a language dependency into the detection task, where artifacts may be more or less perceptible based on the acoustic properties and the sophistication of the synthesis for a given language. Consequently, while the ultimate goal might be language-agnostic countermeasures, the current landscape necessitates consideration of language-specific nuances, especially concerning the maturity and quality of available speech synthesis technologies.

There were certain limitations arising from the interplay of language and synthesis quality in this study. First, the observed language-dependent performance, exemplified by the challenges in Mandarin Chinese deepfake detection, indicates that the effectiveness of our models is not entirely decoupled from linguistic factors. This is likely influenced by the varying quality of generated speech across languages present in our datasets, which complicates the pursuit of purely language-agnostic countermeasures. Second, while the JMDS dataset is multilingual, the uneven distribution of utterances across different languages could impact the training and evaluation of truly balanced, language-agnostic models. Third, our cross-dataset evaluation was constrained by the availability of consistent labeling across external benchmarks, preventing direct evaluation on datasets like ASVspoof 2024. Finally, we utilized relatively general hyperparameter settings, and more targeted optimization for specific datasets could potentially yield improved results. Future research will aim to mitigate these limitations by enhancing the linguistic balance within JMDS, investigating self-supervised learning for better cross-dataset generalization, and exploring more refined, dataset-specific hyperparameter 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 curated 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 Section III. 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 publicly available resources.

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

This work was partially supported by JSPS KAKENHI (25K21245, 23K18491, and 25H01139) and JST Program for co-creating startup ecosystem (JPMJSF2318). 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.

References

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

[2] Z. Wu, N. W. D. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. W. D. Evans, M. Sahidullah, V. Vestman, T. H. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, and Z. Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, p. 101114, 2020.

[4] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, September 18-22, 2022*. Incheon, Korea: ISCA, 2022, pp. 2278–2282.

[5] B. Li, Y. Zhang, T. N. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, May 12-17, 2019*. Brighton, United Kingdom: IEEE, 2019, pp. 5621–5625.

[6] 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

[7] Z. Ba, Q. Wen, P. Cheng, Y. Wang, F. Lin, L. Lu, and Z. Liu, "Transferring audio deepfake detection capability across languages," in *Proceedings of the ACM Web Conference 2023*, ser. WWW '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2033–2044.

[8] 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. of INTERSPEECH 2015, September 6-10, 2015*. Dresden, Germany: ISCA, 2015, pp. 2037–2041.

[9] H. Delgado, M. Todisco, M. Sahidullah, N. W. D. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018: The Speaker and Language Recognition Workshop, 26-29 June 2018*. Les Sables d'Olonne, France: ISCA, 2018, pp. 296–303.

[10] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019, October 10-12, 2019*. Timisoara, Romania: IEEE, 2019, pp. 1–10.

[11] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," in *Proc. of NeurIPS Track on Datasets and Benchmarks 2021, December 2021, virtual*, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract-round2.html

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[13] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "ADD 2022: the first audio deep synthesis detection challenge," in *Proc. of ICASSP 2022, 23-27 May 2022*. Virtual and Singapore: IEEE, 2022, pp. 9216–9220.

[14] 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 Communication*, vol. 164, p. 103122, 2024.

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

[16] 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, S. Liang, Z. Lian, S. Nie, and H. Li, "ADD 2023: the second audio deepfake detection challenge," in *Proceedings of the Workshop on Deepfake Audio Detection and Analysis co-located with 32th International Joint Conference on Artificial Intelligence (IJCAI 2023), August 19, 2023*, ser. CEUR Workshop Proceedings, vol. 3597. Macao, China: CEUR-WS.org, 2023, pp. 125–130. [Online]. Available: https://ceur-ws.org/Vol-3597/paper21.pdf

[17] J. J. Bird and A. Lotfi, "Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion," *CoRR*, vol. abs/2308.12734, 2023.

[18] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," 2024. [Online]. Available: https://arxiv.org/pdf/2401.09512v5

[19] 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. of EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. ACL, 2024, pp. 4977–4983. [Online]. Available: https://aclanthology.org/2024.emnlp-main.286

[20] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, M. Jeong, G. Zhu, Y. Zang, Y. Zhang, S. Maiti, F. Lux, N. Müller, W. Zhang, C. Sun, S. Hou, S. Lyu, S. L. Maguer, C. Gong, H. Guo, L. Chen, and V. Singh, "ASVspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech," 2025. [Online]. Available: https://arxiv.org/abs/2502.08857

[21] 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 *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 921–928.

[22] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "SafeEar: Content Privacy-Preserving Audio Deepfake Detection," in *Proc. of ACM SIGSAC 2024*, ser. CCS '24. New York, NY, USA: ACM, 2024, p. 3585–3599. [Online]. Available: https://doi.org/10.1145/3658644.3670285

[23] T. Kirill, P. Cummer, P. Pherwani, J. Aslam, M. Davinroy, P. Bautista, L. Cassani, and M. Stamm, "Safe: Synthetic audio forensics evaluation challenge," in *Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSEC '25. New York, NY, USA: ACM, 2025, p. 174–180.

[24] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[25] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *CoRR*, vol. abs/1711.00354, 2017. [Online]. Available: http://arxiv.org/abs/1711.00354

[26] İmdat Celeste, "The M-AILABS Speech Dataset," https://github.com/imdatceleste/m-ailabs-dataset, 2020, accessed: April 2025.

[27] 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. of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[28] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. W. D. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2507–2522, 2023.

[29] 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.

[30] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. of INTERSPEECH 2020, October 25-29*. Virtual Event, Shanghai, China: ISCA, 2020, pp. 2757–2761.

[31] 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 *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017*. South Korea: IEEE, 11 2017, pp. 1–5.

[32] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus," in *Interspeech 2021*. Czechia: ISCA, 2021, pp. 2756–2760.

[33] 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 *Interspeech 2021*. Czechia: ISCA, 2021, pp. 3665–3669.

[34] C. O. Mawalim, K. Galajit, D. P. Lestari, W. P. Pa, and M. Unoki, "Challenges in Speech Spoofing Countermeasures for Southeast Asian Languages," ASJ Spring Meeting 2025, Saitama, Japan, 2025.

[35] 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. of (iSAI-NLP)*. Thai: IEEE, 2023, pp. 1–6.

[36] S. A. Arief, C. O. Mawalim, and D. P. Lestari, "Indonesian Speech Anti-Spoofing System: Data Creation and Convolutional Neural Network Models," in *Proc. of ICAICTA 2024*. Singapore: IEEE, 2024, pp. 1–6.

[37] C. O. Mawalim, S. A. Arief, and D. P. Lestari, "InaSAS: Benchmarking Indonesian Speech Antispoofing Systems," *APSIPA Transactions on Signal and Information Processing*, vol. 14, no. 3, pp. –, 2025. [Online]. Available: http://dx.doi.org/10.1561/116.20240080

[38] V. Hoang, V. Pham, H. Xuan, P. Nhi, P. Dat, and T. Nguyen, "VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification," in *Proc. Interspeech 2024*. Kos Island, Greece: ISCA, 2024.

[39] 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. of OCOCOSDA 2024, October 17-19, 2024*. Hsinchu City, Taiwan: IEEE, 2024, pp. 1–5.

[40] R. Kolobov, O. Okhapkina, O. Omelchishina, A. Platunov, R. Bedyakin, V. Moshkin, D. Menshikov, and N. Mikhaylovskiy, "MediaSpeech: Multilanguage ASR Benchmark and Dataset," 2021. [Online]. Available: https://arxiv.org/abs/2103.16193

[41] T. Javed, J. Nawale, E. George, S. Joshi, K. Bhogale, D. Mehendale, I. Sethi, A. Ananthanarayanan, H. Faquih, P. Palit, S. Ravishankar, S. Sukumaran, T. Panchagnula, S. Murali, K. Gandhi, A. R, M. M, C. Vaijayanthi, K. Karunganni, P. Kumar, and M. Khapra, "IndicVoices: Towards building an Inclusive Multilingual Speech Dataset for Indian Languages," in *Findings of the ACL 2024*. Bangkok, Thailand: ACL, Aug. 2024, pp. 10 740–10 782.

[42] A. Adila, C. O. Mawalim, and M. Unoki, "Detecting Spoof Voices in Asian Non-Native Speech: An Indonesian and Thai Case Study," in *Proc. of APSIPA ASC 2024, December 3-6*. Macau: IEEE, 2024, pp. 1–6.

[43] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. of (ICASSP) 2021*, 2021, pp. 6493–6497.

[44] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Proc. of Interspeech 2019, Graz, Austria, September 15-19*. ISCA, 2019, pp. 1078–1082.

[45] J. Jung, H. Heo, H. Tak, H. Shim, J. S. Chung, B. Lee, H. Yu, and N. W. D. Evans, "AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, 23-27 May 2022*. Virtual and Singapore: IEEE, 2022, pp. 6367–6371.

[46] J. Brown, "Calculation of a Constant Q Spectral Transform," *Journal of the Acoustical Society of America*, vol. 89, pp. 425–, 01 1991.

[47] 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, "ASVspoof 5 Evaluation Plan," ASVspoof consortium, Tech. Rep., 2024. [Online]. Available: http://www.asvspoof.org/

[48] C. O. Mawalim, Y. Wang, A. Adila, S. Okada, and M. Unoki, "Robust multilingual audio deepfake detection through hybrid modeling," in *Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSEC '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 181–192. [Online]. Available: https://doi.org/10.1145/3733102.3736706

[49] 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. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

Appendix
Details related to the JMDS dataset

TABLE VIII: 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**
English (eng)	3.21 ± 0.26	2.91 ± 0.65
Mandarin (zho)	3.19 ± 0.20	2.38 ± 0.63
Hindi (hin)	2.81 ± 0.29	3.14 ± 0.4
Italian (ita)	3.17 ± 0.24	3.04 ± 0.47
Arabic (arb)	2.88 ± 0.33	3.29 ± 0.31
Spanish (spa)	3.14 ± 0.28	3.15 ± 0.42
Polish (pol)	3.24 ± 0.24	3.15 ± 0.42
German (deu)	3.12 ± 0.31	3.09 ± 0.42
French (fra)	3.26 ± 0.21	3.18 ± 0.38
Russian (rus)	3.29 ± 0.19	3.16 ± 0.41
Portuguese (por)	–	3.22 ± 0.4
Japanese (jpn)	3.09 ± 0.35	3.22 ± 0.45
Ukrainian (ukr)	3.22 ± 0.22	2.92 ± 0.52
Vietnamese (vie)	3.08 ± 0.33	3.3 ± 0.21
Thai (tha)	3.17 ± 0.25	3.08 ± 0.44
Indonesian (ind)	2.95 ± 0.38	3.12 ± 0.38
Burmese (mya)	3.16 ± 0.26	3.14 ± 0.3
All	**3.14 ± 0.3**	**2.93 ± 0.61**

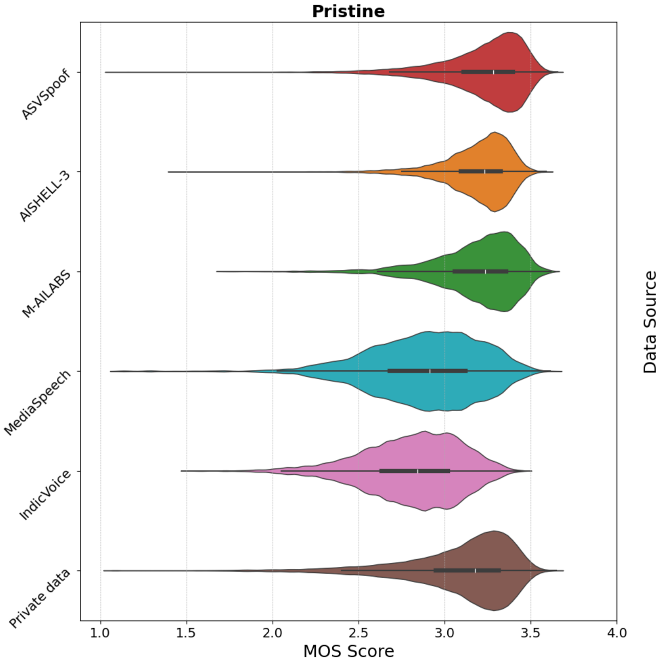

Fig. 7: Violin plots illustrating the dDistribution of MOS scores for pristine speech across different data sources in the JMDS dataset. Pristine sources generally show consistently high perceptual quality.

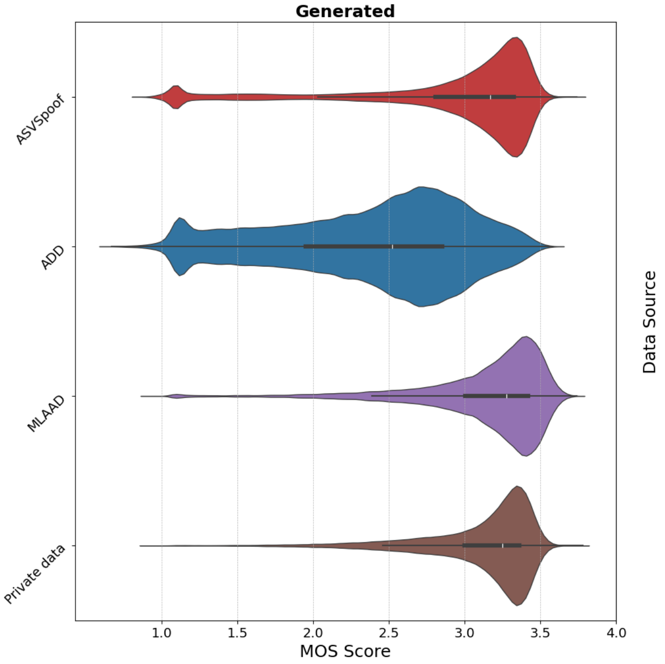

Fig. 8: Violin plots illustrating the dDistribution of MOS scores for generated speech across different data sources in the JMDS dataset. Among the generated sources, only ADD exhibits a notably broad range of audio quality, indicating varied synthesis characteristics, while others like ASVspoof, MLAAD, and Private data tend to produce more consistent high-MOS outputs.

TABLE IX: 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

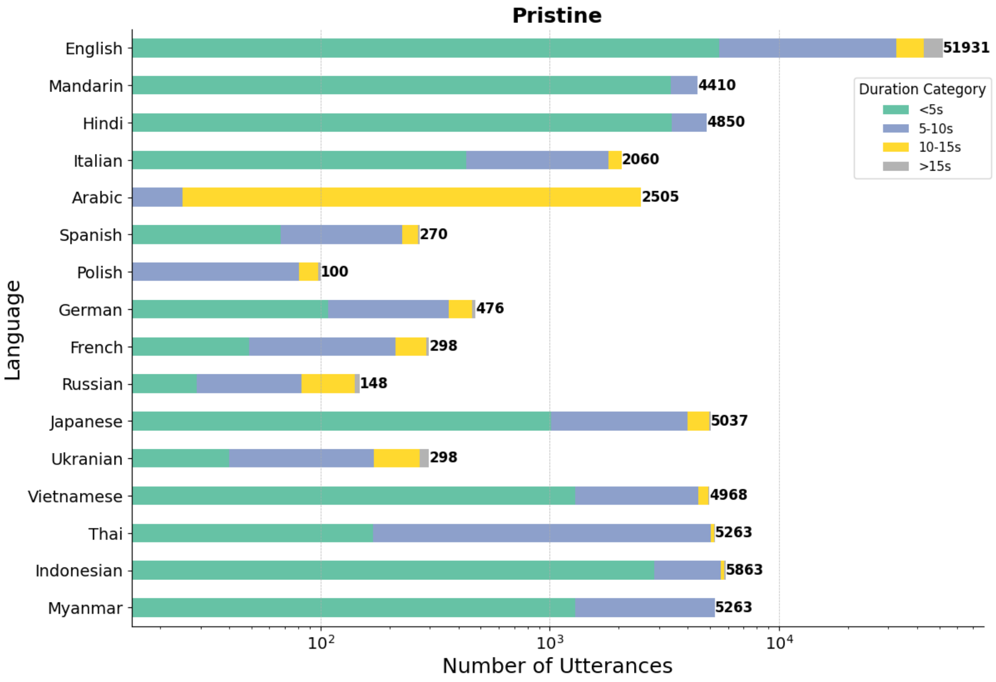

Fig. 9: A detailed breakdown of utterance duration distribution across different languages for pristine samples in the JMDS dataset, illustrated on a logarithmic scale.

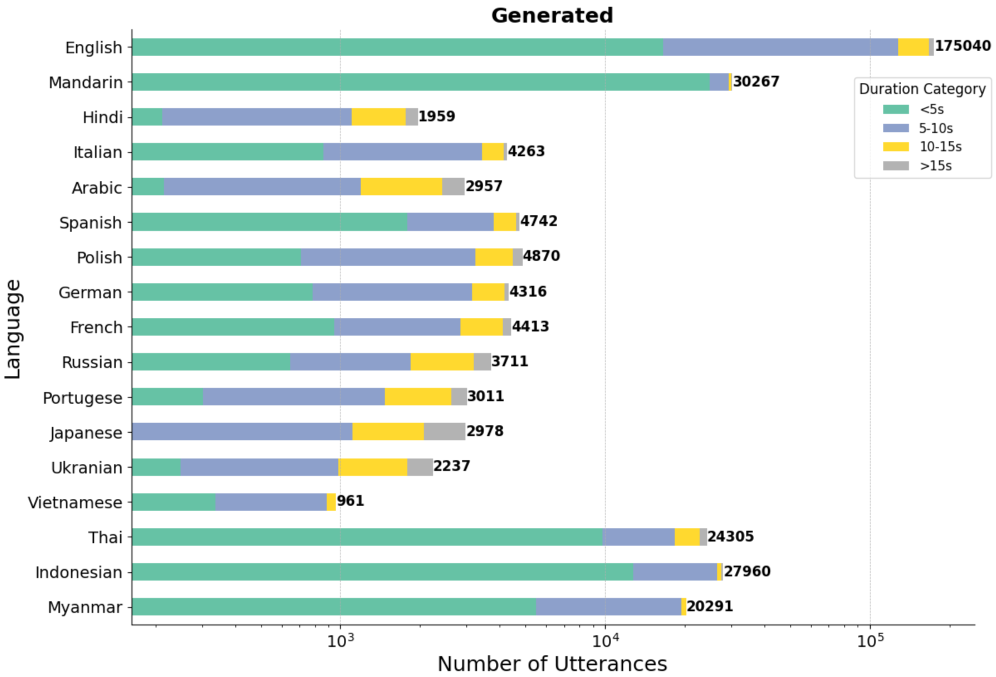

Fig. 10: A detailed breakdown of utterance duration distribution across different languages for generated samples in the JMDS dataset, illustrated on a logarithmic scale.

Biography Section

CANDY OLIVIA MAWALIM (Member, IEEE) received her B.S. in Computer Science from Institut Teknologi Bandung (ITB), Indonesia, and M.S. and Ph.D. from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), with the Ph.D. being awarded in 2022. She was selected as 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 at the School of Information Science at JAIST, where she currently serves as 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 his B.S. in Computer Science and Technology from Dalian University of Foreign Languages in Dalian, China, and his M.S. from the 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 her B.S. in Computer Science from Institut Teknologi Bandung (ITB), Indonesia, and her M.S. in Speech Processing from the School of Information Science at the 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 machine learning and deep learning methods.

SHOGO OKADA (Member, IEEE) directs the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan and is a professor at JAIST. He obtained his Ph.D. in 2008 from Tokyo Institute of Technology in Japan. He joined Kyoto University as a project assistant professor in 2008, Tokyo Institute of Technology, as a tenured assistant professor in 2011. He joined IDIAP Research Institute in 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 the IEEE and ACM.

MASASHI UNOKI (Member, IEEE) received his M.S. and Ph.D. in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His research interests include auditory-motivated signal processing and the modeling of auditory systems. 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 at 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. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of 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 the ASJ in 1999, 2010, and 2013, as well as the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation in 2005.