Bridging the SEA Gap: An Initial Benchmark for Neural Audio Codec-Synthesized Speech Deepfakes in South-East Asian Languages

Orchid Chetia Phukan[*1] , **Girish**[1,2] , **Mohd Mujtaba Akhtar**[1,3] , **Arun Balaji Buduru**[1]
[1]IIIT-Delhi, India
[2]UPES, India
[3]VBSPU, India
orchidp@iiitd.ac.in

Abstract

Codecfakes (CFs) are a type of speech deepfakes generated through Audio Language Models (ALMs), with Neural Audio Codecs (NACs) forming the core mechanism for speech encoding and generation. CFs exhibit distributional characteristics that differ from vocoder-based deepfakes, causing detectors trained on vocoder data to generalize poorly to CFs detection. Although this has led to the development of CF detection benchmarks, existing resources are largely confined to English—and to a limited extent Chinese—leaving South-East Asian (SEA) languages unexplored. To bridge this gap, we introduce SEA-CF, the first large-scale benchmark for CF detection spanning multiple SEA languages, diverse speaker profiles, and a wide range of NAC architectures. SEA-CF is constructed by synthesizing publicly available real speech corpora. Our experiments show that state-of-the-art (SOTA) CF detectors trained on English-centric datasets fail to generalize to SEA speech due to language-specific phonetic structures, tonal variations, and rich prosodic diversity. We further conduct a comprehensive zero-shot and fine-tuned evaluation of recent SOTA ALMs on SEA-CF. Fine-tuning the ALMs improves performance, however, these are very large being impractical for real-world application due to their scale, particularly in low-resource and latency-constrained settings. To address this limitation, we propose a novel small-ALM, **GARUDA** tailored for CF detection, which delivers strong performance while remaining lightweight. Extensive evaluations demonstrate that the proposed Small-ALM outperforms strong end-to-end and ALM-based baselines, establishing a new, practical direction for robust CF detection in SEA languages and beyond.

1 Introduction

Recent advances in generative speech technologies, including text-to-speech (TTS) and voice conversion (VC), have enabled the synthesis of speech that closely matches natural human speech. In many cases, distinguishing machine-generated speech from real speech has become increasingly challenging, as modern systems can produce perceptually realistic voices across a wide range of languages. While these developments unlock significant opportunities in applications such as customer-facing services, healthcare, and entertainment, they also pose substantial security and societal risks. These include impersonation attacks, deepfake-enabled misinformation, and voice-based fraud in high-stakes domains such as digital banking and e-governance. To mitigate these threats, the speech research community has introduced a range of benchmarks for building models aimed at identifying synthetic speech [Wu *et al.*, 2015; Nautsch *et al.*, 2021; Liu *et al.*, 2023; Wang *et al.*, 2024]. Detection approaches have evolved steadily over time, showing consistent improvements over earlier methods. Early efforts relied on handcrafted acoustic features combined with classical machine learning techniques [Patel and Patil, 2015]. This was later succeeded by deep learning–based models [Scardapane *et al.*, 2017; Gomez-Alanis *et al.*, 2019; Chintha *et al.*, 2020; Jung *et al.*, 2022; Müller *et al.*, 2024]. Further, from beginning of the current decade, a substantial leap in speech deepfake detection performance was driven by the emergence of large pre-trained models (PTMs) [Wang and Yamagishi, 2022; Tak *et al.*, 2022]. Consequently, the research community has extensively explored PTMs such as Whisper, XLS-R, Wav2vec2, WavLM etc. and proposing a wide range of novel detection frameworks that leverage these PTMs [Kawa *et al.*, 2023; Phukan *et al.*, 2024; Huang *et al.*, 2025].,More recently, the community has began the usage of Audio Language Models (ALMs) such as Qwen2-Audio for speech deepfake detection. They have evaluated these ALMs either in a zero-shot or fine-tuning them for improved downstream performance [Gu *et al.*, 2025; Chuchra *et al.*, 2026].

However, the majority of existing studies focus primarily on speech deepfakes generated using vocoder-based approaches, including traditional TTS and VC systems. This emphasis overlooks a rapidly emerging class of speech deepfakes i.e. Codecfakes (CFs) produced by ALMs, which differ fundamentally in their generation mechanisms and acoustic characteristics. ALMs rely on Neural Audio Codecs (NACs), including SoundStream, EnCodec, and DAC, as the backbone for speech

*Corresponding Author

encoding and generation. For instance, AudioLM employs SoundStream for discrete audio tokenization and synthesis [Borsos *et al.*, 2023]. In response to this emerging threat, [Wu *et al.*, 2024] and [Lu *et al.*, 2024b] took the first steps by introducing dedicated CF benchmarks for the research community. Their studies demonstrated that state-of-the-art (SOTA) models trained on vocoder-based speech deepfake datasets fail to generalize when evaluated on CFs, highlighting the necessity of CF-specific detection approaches. These benchmarks have since fostered meaningful progress in CF detection [Zhu *et al.*, 2025]; however, they remain largely English-centric. While [Lu *et al.*, 2024b] additionally includes Chinese as a secondary language, other languages remain entirely unexplored. This leaves a substantial gap in defense against CFs as previous research have shown that cross-lingual speech deepfake detection performance generally decreases [Ba *et al.*, 2023; Phukan *et al.*, 2024].

In this work, we focus on CF detection in South-East Asian (SEA) languages. Our motivation stems from the fact that SEA, home to over 700 million people, is among the most linguistically and culturally diverse regions worldwide. Countries such as Indonesia, Thailand, Vietnam, the Philippines, Malaysia, Myanmar, Cambodia, Laos, and Singapore collectively encompasses languages spanning multiple families, including Austronesian (e.g., Malay, Indonesian, Tagalog), Austroasiatic (e.g., Vietnamese, Khmer), Tai–Kadai (e.g., Thai, Lao), and Sino-Tibetan (e.g., Burmese), each characterized by distinct phonetic, tonal, and prosodic structures. While [Wu *et al.*, 2025] introduced the first benchmark for speech deepfake detection in SEA languages, their study is limited to vocoder-based generation and does not consider CFs; moreover, the dataset has not been publicly released. To address these limitations, we introduce SEA-CF, the first publicly available and large-scale benchmark dedicated to CF detection in SEA languages, covering multiple languages, diverse speaker profiles, and a broad range of SOTA NAC architectures. SEA-CF is constructed by synthesizing publicly available real speech corpora using multiple NACs. Experimental results show that SOTA CF detectors trained on English-centric datasets fail to generalize to SEA speech, primarily due to linguistic differences. However, when jointly training previous benchmark CF dataset with SEA-CF, the performance improves, this shows the requirement of domain-specific training. We further perform comprehensive zero-shot and fine-tuned evaluations of recent SOTA ALMs. We evaluate these ALMs as previous research has shown their effectiveness compared to traditional end-to-end or PTM-based approaches for detecting speech deepfakes generated through vocoder-based approaches [Gu *et al.*, 2025; Chuchra *et al.*, 2026]. While fine-tuning improves detection performance, the large model sizes of these ALMs make them impractical for real-world applications, particularly in low-resource and latency-sensitive scenarios. Such real-world application is need of the hour for deepfake detection tools to be employed in real-world detection scenarios. To overcome this limitation, we propose a novel Small-ALM, **GARUDA**, tailored for CF detection. Extensive evaluations demonstrate that the proposed **GARUDA** consistently outperforms strong end-to-end and ALM-based baselines across both SEA-CF and previous CF detection benchmarks (achieving SOTA) while being lightweight and lesser inference time. **The main contributions are summarized as follows**: (i) We present SEA-CF, the first publicly available, large-scale benchmark for CF detection in SEA languages, encompassing multiple SEA languages, diverse speaker profiles, and a wide range of SOTA NACs. (ii) Extensive experiments on SEA-CF reveal that SOTA CF detectors trained on English-centric data generalize poorly to SEA languages. In contrast, joint training with SEA-CF and existing CF benchmarks improves performance, underscoring the necessity of in-domain training. (iii) We further perform comprehensive zero-shot and fine-tuned evaluations of recent SOTA ALMs. Our results reveal performance gains from fine-tuning. (iv) To address the real-world applicability challenges with large ALMs, we propose **GARUDA**, a novel Small ALM tailored for CF detection. Extensive evaluations show that **GARUDA** achieves SOTA performance across both SEA-CF and existing CF detection benchmarks while remaining lightweight and offering significantly reduced inference time. *The SEA-CF dataset, code, and evaluation resources are available at:* [1]

This work supports SDG 16 (Peace, Justice, and Strong Institutions) by strengthening defenses against codec-based speech deepfakes, especially for underrepresented SEA languages, and SDG 10 by reducing linguistic and regional gaps in deepfake detection. It also aligns with the "Leave No One Behind" principle by improving access to practical detection resources for low-resource language communities. The work was informed by academic collaboration and domain feedback from cybersecurity-oriented stakeholders, which helped shape the focus on lightweight and deployable CF detection systems. In particular, co-author Dr. Arun Balaji Buduru, a faculty member at IIIT-Delhi who is also associated with cybersecurity projects and fraud detection initiatives, provided valuable insights into the practical challenges faced in real-world deployments. These discussions highlighted several key requirements that directly influenced the design of our framework: (i) *Language coverage gap*: Existing fraud detection systems are predominantly English-centric, leaving users of SEA languages vulnerable to emerging audio fraud threats. This motivated the development of the SEA-CF dataset to improve linguistic coverage and inclusivity. (ii) *Low-latency deployment requirements*: Real-world fraud detection systems must operate under stringent latency and resource constraints. Consequently, we focused on a lightweight audio language model (less than 1B parameters) and optimized the framework to achieve an inference time of 1.21 seconds. (iii) *Robustness to unseen NACs*: Fraudulent audio encountered in practice may originate from previously unseen synthesis pipelines. This motivated our explicit seen-versus-unseen codec evaluation protocol to assess cross-codec generalization and deployment robustness.

2 Related Works

[Wu *et al.*, 2024] and [Lu *et al.*, 2024b] made the research community to focus on the need for building models for CF detection as models trained on previous TTS and VC synthesis-based datasets fail here. [Wu *et al.*, 2024] curated the CF

[1] https://helixometry.github.io/SEACodecFake/.

dataset using the English VCTK corpus, while, [Lu *et al.*, 2024b] developed CF corpora covering English and Chinese by leveraging VCTK and AISHELL-3 as source corpora. Both of these studies leverage AASIST-based modeling for CF detection. Subsequent work further expanded the CF detection landscape by incorporating a broader range of NAC families while continuing to rely on AASIST-based architectures as well as proposing an alternative CF detection strategy based on sharpness-aware minimization to improve robustness [Xie *et al.*, 2025]. Despite these advances, existing CF datasets remain almost exclusively focused on high-resource languages, primarily English and Chinese, highlighting a clear lack of linguistic diversity. Notably, no prior work has systematically investigated CF detection in SEA languages. While [Cui *et al.*, 2025] was the first to extend CF detection beyond English and Chinese by introducing a multilingual CF dataset, SEA languages were not included. In addition, [Wu *et al.*, 2025] proposed a dataset to facilitate speech deepfake detection in SEA languages; however, their work is limited to vocoder-based generation methods such as TTS and VC, and the dataset has not yet been made publicly available. Recent work has also begun extending synthetic speech forensics beyond closed-set binary detection, including Indic-language codec-fake detection [Girish *et al.*, 2026], cross-lingual synthetic-speech attribution and emotional manipulation analysis [Girish *et al.*, 2025], open-set generator attribution for unseen synthesizers [Akhtar *et al.*, 2026a], geometry-aware generalization across diverse speech synthesis paradigms [Sheth *et al.*, 2025], and codec-fake detection in high-stakes healthcare settings [Akhtar *et al.*, 2026b]. These studies further highlight the need for codec-specific, language-diverse, and efficient detection benchmarks. To bridge these gaps, we introduce the first large-scale and comprehensive benchmark dedicated to CF detection in SEA languages. Furthermore, we propose **GARUDA**, a novel Small-ALM designed for CF detection in SEA languages and beyond.

3 SEA-CF

In this section, we first describe the real-speech source corpora used to construct SEA-CF. We then present an overview of the SOTA NACs employed for CF generation, followed by a detailed description of the end-to-end data generation pipeline used to build the SEA-CF benchmark.

Real-speech sources: We adopt the same set of languages as SEA-Spoof [Wu *et al.*, 2025], namely Tamil, Hindi, Thai, Indonesian, Malay, and Vietnamese. For real-speech sources, we largely follow the datasets used in SEA-Spoof, including Mozilla Common Voice [Ardila *et al.*, 2020] for multilingual coverage, GigaSpeech2 [Yang *et al.*, 2025] for Indonesian, the Thai Dialect Corpus [Suwanbandit *et al.*, 2023] for Thai, and the VIVOS corpus [Luong and Vu, 2016] for Vietnamese. For Malay, we combine the Conversational Malay Speech Corpus[2] with a curated Malaysian YouTube dataset processed using Whisper-Large[3]. In contrast to SEA-Spoof, for Tamil and Hindi, we utilize the Indic-SUPERB corpus [Javed *et al.*, 2023] , one of the largest publicly available datasets for these languages. Indic-SUPERB has also been employed in prior studies on speech deepfake detection for Tamil and Hindi, primarily focusing on vocoder-based TTS and VC generation approaches [Sharma *et al.*, 2025].

Neural Audio Codecs (NACs): We adopt widely-used, publicly released NACs that are straightforward to reproduce and have been used by previous CF detection studies for generation for CFs [Lu *et al.*, 2024b; Wu *et al.*, 2024]. **DAC** [Kumar *et al.*, 2024]: We employ DAC with 16 kHz, 24 kHz, and 44 kHz configurations. **Encodec** [Défossez *et al.*, 2022]: We use its 24 kHz and 48 kHz models. **SoundStream** [Zeghidour *et al.*, 2021]: We use its 16 kHz version. **SpeechTokenizer** [Zhang *et al.*, 2024]: We utilize its default 16 kHz configuration. **FunCodec** [Du *et al.*, 2024]: We adopt the official 16 kHz version. **AudioDec** [Wu *et al.*, 2023]: We incorporate AudioDec with 28 kHz and 48 kHz variants. **SNAC** [Siuzdak *et al.*, 2024]: We use SNAC operating at 24 kHz, 32 kHz, and 44 kHz. **MIMI** [Défossez *et al.*, 2024]: We utilize MIMI which operates at 24 kHz. We curate a repository compiling all the NACs used in our study and a easy to use pipeline for using them [4].

Generation Pipeline of SEA-CF: To build SEA-CF, we employ a controlled resynthesis pipeline inspired by prior work on CF generation [Wu *et al.*, 2024]. CF samples are created by encoding and reconstructing real speech drawn from the SEA-CF source corpora using a diverse collection of NACs given above. Given an input speech waveform s, a NAC encoder $\mathcal{F}$ converts the signal into a discrete latent representation $q = \mathcal{F}(s)$. This representation is then passed through the corresponding decoder $\mathcal{G}$ to generate a reconstructed waveform $\hat{s} = \mathcal{G}(q)$, which is treated as the associated CF sample. This encoding–decoding process preserves the underlying linguistic content and speaker characteristics while introducing codec-specific artifacts unique to each NAC. Applying this procedure to every utterance across all selected NACs yields parallel datasets in which each real speech sample has a one-to-one CF counterpart for every codec configuration. To evaluate detection robustness, we define two experimental settings. (i) Seen setting: both training and evaluation involve CF samples generated using the same set of NACs (e.g., SNAC variants, DAC variants, Encodec, SoundStream, and SpeechTokenizer). (ii) Unseen setting: evaluation is conducted on CF samples synthesized using NACs not encountered during training (e.g., FunCodec variants, AudioDec variants, and MIMI), enabling assessment of cross-codec generalization. This pipeline is followed consistently across all SEA languages included in SEA-CF. We next describe the training, validation, and testing splits used in SEA-CF. For the (ii) Unseen evaluation setting, the test set exclusively consists of CF samples generated using NACs that are not observed during training. **Hindi and Tamil:** Since IndicSUPERB provides predefined training, validation, and test splits, we follow these splits directly for CF generation. IndicSUPERB further includes two test partitions, namely test-known and test-unknown. We use the test-known split for the (i) Seen evaluation setting and the test-unknown split for the (ii) Unseen evaluation setting. **All Other Languages:** For

[2]https://magichub.com/datasets/malay-conversational-speech-corpus/

[3]https://huggingface.co/mesolitica/datasets

[4]https://github.com/CodeVault-girish/Neural-Codecs

the remaining SEA languages, we follow the protocol in [Wu *et al.*, 2025], as the same real-speech sources are adopted. We randomly partition the data into training, validation, and testing sets using an 8:1:1 ratio. This partitioning is strictly preserved during CF generation to prevent data leakage.

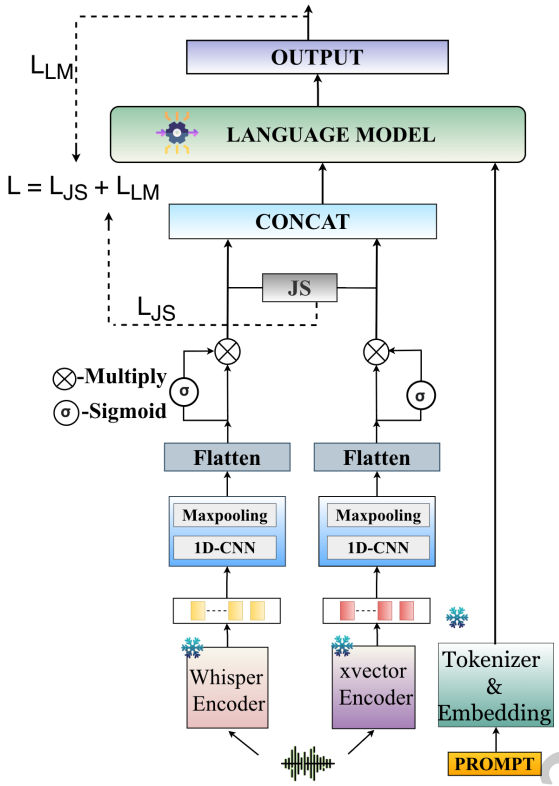

Figure 1: Proposed Framework: **GARUDA**

4 Methodology

In this section, we introduce our proposed framework, **GARUDA**. We first provide an overview of the overall architecture of **GARUDA**, followed by a detailed description of the audio encoders employed. Finally, we explain the language model (LM)-based decoder that enables high-level reasoning over fused audio representations. **GARUDA** adopts a dual-encoder design comprising Whisper and x-vector models to capture complementary aspects of speech. Whisper is leveraged to encode semantic and linguistic information, benefiting from its large-scale speech–text pretraining, while the x-vector encoder is employed to capture prosodic cues, such as pitch, tone attributing to its speaker recognition pretraining. Such dual-encoder architectures have been shown to be effective for speech deepfake detection in prior work, particularly through the fusion of semantic and prosodic cues [Phukan *et al.*, 2024]. Similar dual-encoder fusion strategies have also been adopted in large ALMs for speech deepfake detection [Ranjan *et al.*, 2025]. However, such approaches have not yet been explored in the context of CF detection within ALM-based frameworks. Furthermore, formulating speech deepfake detection as an audio question-answering task using ALMs has been shown to be effective in prior studies [Gu *et al.*, 2025; Li *et al.*, 2025], as it enables the model to perform context-aware reasoning and leverage the generative capabilities of ALMs beyond conventional discriminative classification.

Audio Encoders: **Whisper** [Radford *et al.*, 2023]: It is a transformer-based encoder–decoder model trained in a multilingual, multi-task setting across 96 languages. In our framework, we utilize only the frozen encoder component to extract speech representations, producing a 512-dimensional embedding via average pooling. Owing to its training objective centered on automatic speech recognition, Whisper is particularly effective at encoding semantic and linguistic content in speech. We use the base version in our study which consists of 74M parameters. **x-vector** [Snyder *et al.*, 2018]: It is a time-delay neural network trained in a supervised manner for speaker recognition. The model is trained on a combination of the VoxCeleb1 and VoxCeleb2 datasets and contains approximately 4.2 million parameters. We use the frozen model and extract 512-dimension representations through average pooling.

Language Model Decoder: We adopt Qwen2-0.5B as the pretrained decoder LM backbone in our framework [Team and others, 2024]. The use of a small LM (Less than 1B parameters) aligns with our motivation of designing a lightweight small-ALM that enables effective audio reasoning without the computational overhead of large-scale LLMs. This is due to the fact that the decoder LM generally remains heavyweight compared to its audio encoder counterparts in an ALM framework. Lighter the decoder LM lighter the ALM. As the smallest model in the Qwen2 family, Qwen2-0.5B offers an efficient yet capable architecture. Qwen2 LM family has demonstrated strong performance across speech and audio–language modeling tasks, including speech deepfake detection [Li *et al.*, 2025] and speech emotion recognition [Su *et al.*, 2025].

Working Flow of GARUDA: The archiecture of **GARUDA** is shown in Figure 1. Given a speech utterance, the objective of **GARUDA** is to determine whether the input audio is genuine or manipulated by generating a concise natural-language decision. To this end, we extract two complementary representations from the input: a semantic representation using Whisper and a prosodic representation using x-vector. These representations are subsequently processed through a convolutional module consisting of a 1D-CNN layer with a kernel size of three, followed by max pooling. After flattening, the features are passed through a sigmoid gating module that selectively filters the input representations, allowing only important information from the previous stage to passed through to the next layer. Suppose x and y represent the two feature vectors obtained from the two audio encoder branches. To encourage information-level consistency between these heterogeneous representations prior to fusion, we employ Jensen–Shannon (JS) divergence [Sutter *et al.*, 2020] as a novel loss function. Since JS divergence is defined over probability distributions, we first transform the feature vectors into feature-wise distributions using temperature-scaled softmax normalization: $p_x = \text{softmax}\left(\frac{x}{\tau}\right)$ and $p_y = \text{softmax}\left(\frac{y}{\tau}\right)$ where τ denotes a temperature parameter controlling the smoothness of the distributions. We then define the mix-

ture distribution as: $m = \frac{1}{2}(p_x + p_y)$. The JS alignment loss is computed as: $\mathcal{L}_{\text{JS}} = \frac{1}{2}\,\text{KL}(p_x \| m) + \frac{1}{2}\,\text{KL}(p_y \| m)$ where the Kullback–Leibler (KL) divergence is defined as: $\text{KL}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$. For the JS loss, lower the value shows greater alignment and we will in later stage jointly optimize it with the language modeling loss. Following alignment, the two representations are concatenated and passed through a fully connected network of 90 neurons. The concatenated output is then projected into the LM embedding space and injected as a continuous prefix to the decoder LM. We use the prompt: *Is the speech sample fake or real? Reply in one word "fake" or "real"*. Here the model is trained to predict only the word *"fake"* or *"real"* as the output. As our study doesn't focus on seeing the variance in performance due to the usage of different prompts, so we only experiment with a single prompt template. Finally, the decoder is trained using the standard language modeling objective $\mathcal{L}_{\text{LM}}$. The language modeling objective minimizes the negative log-likelihood of the target sequence given a predicted sequence. The final training objective is given by: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda\,\mathcal{L}_{\text{JS}}$ where λ controls the contribution of the JS alignment loss.

5 Experiments

5.1 Training Details

GARUDA is trained in a supervised setting on the combination of previous benchmark CF dataset [Lu *et al.*, 2024a] which contains English, Chinese and SEA-CF. We use [Lu *et al.*, 2024a] dataset as English and Chinese both are spoken in Singapore. We train on the combine training split as [Lu *et al.*, 2024a] have dedicated training, validation and testing split. Training of **GARUDA** is carried out in two formats: (i) Only training the projection module which consists of convolutional blocks, JS alignment loss, and the fully connected network (ii) training both the projection module as well as the LM decoder with LoRA. We use LoRA as fine-tuning full LM is cost intensive and LoRA provides a efficient way to fine-tune LMs. For (i), we train the models for 5 epochs with a batch size of 32 and learning rate of 1e-4. While for (ii), we train the models for 3 epochs with same batch size and learning rate of 1e-5. For LoRA, we set the rank to 8 and the scaling factor to 32. LoRA adapters are inserted into the query and value projection layers of the LM. For both the training formats, we apply dropout to prevent overfitting as well as use the same optimizer AdamW. We also kept the τ and λ values to 0.5 and 0.4 respectively after initial experimentation on the validation set. For our experiments, we use A100 GPUs.

5.2 Experimental Results

For testing, we test on Codecfake [Lu *et al.*, 2024a] and SEA-CF individually to quantify the performance. We use Accuracy (ACC) and Equal Error Rate (EER) as the evaluation metrics as used by previous speech deepfake detection research including CF detection studies [Wu *et al.*, 2024; Lu *et al.*, 2024a; Phukan *et al.*, 2024] (Table 1).

SOTA model trained on CodecFake [Lu *et al.*, 2024a] and evaluated on SEA-CF: We first access whether model trained on the prior benchmark [Lu *et al.*, 2024a] which consists of English and Chinese–can generalize to SEA-CF without

Method	SEA-CF		CF [Lu *et al.*, 2024a]		Avg	
	ACC ↑	EER ↓	ACC ↑	EER ↓	ACC ↑	EER ↓
Zero-shot evaluation of ALMs						
Qwen-Audio-Chat	3.72	94.39	14.10	85.80	8.91	90.10
Qwen-Audio-Base	5.41	94.67	16.07	84.36	10.74	89.52
Qwen2-Audio-Chat	5.96	91.71	17.23	81.17	11.60	86.44
Qwen2-Audio-Base	8.41	91.53	19.48	80.47	13.95	86.00
SeaLLMs-Audio-7B	6.23	91.64	18.35	80.25	12.29	85.95
End-to-end method						
AASIST	86.98	15.74	93.09	8.16	90.04	11.95
Pre-Trained Backbone						
Wh-LCCN	87.69	15.22	94.41	7.63	91.05	11.43
Wav2vec2-AASIST	88.71	13.01	95.16	7.08	91.94	10.04
MiO	88.76	12.51	95.64	6.37	92.20	9.44
FT						
SeaLLMs-Audio-7B	88.74	9.64	90.75	6.96	89.75	8.30
Qwen2-Audio-Base	93.88	6.95	95.06	4.21	94.47	5.58
GARUDA						
Only Wh	89.58	11.72	90.12	7.53	89.85	9.63
Only XV	90.99	10.43	93.15	7.28	92.07	8.86
Wh+XV (Concat)	91.56	9.17	94.88	7.16	93.22	8.17
Wh+XV (CA)	93.62	7.04	95.10	6.41	94.36	6.73
Wh+XV (KL)	90.83	9.31	93.46	6.68	92.15	8.00
GARUDA	94.37	6.26	97.00	4.19	95.69	5.23
GARUDA-FT						
Only Wh	91.87	10.31	92.80	6.43	92.34	8.37
Only XV	92.69	9.62	95.31	5.17	94.00	7.40
Wh+XV (Concat)	93.78	8.24	95.40	4.38	94.59	6.31
Wh+XV (CA)	96.07	5.82	97.26	5.49	96.67	5.66
Wh+XV (KL)	94.38	8.11	96.75	4.21	95.57	6.16
GARUDA-FT	**98.41**	**2.78**	**99.36**	**1.68**	**98.89**	**2.23**
Training on Segments of Data						
Qwen2-Audio-Base-FT (75%)	89.51	8.67	93.24	5.92	91.38	7.30
Qwen2-Audio-Base-FT (50%)	86.02	9.48	90.23	7.42	88.12	8.45
Qwen2-Audio-Base-FT (25%)	82.93	10.73	88.07	8.76	85.50	9.75
GARUDA(75%)	92.21	7.48	96.23	5.12	94.22	6.30
GARUDA(50%)	91.48	8.27	95.17	6.56	93.33	7.42
GARUDA(25%)	90.04	9.67	93.82	7.31	91.93	8.49
GARUDA-FT(75%)	95.12	6.34	98.43	4.03	96.77	5.19
GARUDA-FT(50%)	93.56	6.89	98.24	5.47	95.90	6.18
GARUDA-FT(25%)	92.04	8.29	96.26	6.19	94.15	7.24

Table 1: Results on SeaCF and CodecFake (CF) [Lu *et al.*, 2024a]. All results are reported as percentages. Abbreviations: XV: x-vector; Wh: Whisper; FT: fine-tuning. **GARUDA-FT** denotes fine-tuning of the language-model decoder. KL denotes Kullback–Leibler divergence. These scores correspond to evaluation under the seen setting. The same abbreviations are used in Table 2.

Method	SEA-CF		CF[Lu *et al.*, 2024a]		Avg	
	ACC ↑	EER ↓	ACC ↑	EER ↓	ACC ↑	EER ↓
MiO	85.55	13.91	93.44	7.77	88.50	10.84
Qwen2-Audio-Base-FT	92.08	8.15	93.26	5.42	92.67	6.79
GARUDA	92.97	6.88	94.60	5.71	93.79	6.30
GARUDA-FT	**97.11**	**3.17**	**98.06**	**2.23**	**97.59**	**2.70**

Table 2: Performance on Unseen Test Set

any SEA-specific training. We use AASIST as the modeling architecture as it was shown to be effective by both [Lu *et al.*, 2024a] and [Wu *et al.*, 2024] and represents SOTA model. Training and evaluation on CodecFake [Lu *et al.*, 2024a] dataset gives us around 94.08% ACC and 6.76% EER. However, when evaluated on SEA-CF, we get 70.65% ACC 28.13% EER, thus showing extreme degradation in performance and the need for in-domain training on SEA-CF.

Zero-shot evaluation of SOTA ALMs: We evaluate SOTA ALMs in a strictly zero-shot setting by prompting them to classify each utterance with the same prompt as used for **GARUDA**. As shown in Table 1, zero-shot ALMs are generally unreliable for CF detection across both SEA-CF and CodecFake [Lu *et al.*, 2024a]. We consider Qwen-audio and Qwen2-audio families of large ALMs as considered by [Gu *et al.*, 2025] which is first study of evaluating ALMs for speech deepfake detection excluding CF detection. We also include SeaLLMs-Audio-7B [Liu *et al.*, 2025] as this ALM is fine-tuned on SEA languages. The results are presented in Table 1. Qwen2-Audio-Base performs the best across all ALMs. However, the performance across all the ALMs remains quite low. SeaLLMs-Audio-7B despite being trained especially for SEA languages performs lower than Qwen2-Audio-Base. This performance of Qwen2-Audio-Base is also supported by [Gu *et al.*, 2025]. So, we keep Qwen2-Audio-Base as the SOTA ALM architecture for comparison in our succedding experiments.

In-domain Training: Classifier Baselines and SOTA ALMs Finetuning: We present the results with training done on combination of Codecfake [Lu *et al.*, 2024a] and SEA-CF and evaluated on SEA-CF and Codecfake [Lu *et al.*, 2024a]. The results are given in Table 1. We consider several baselines including End-to-end (AASIST [Jung *et al.*, 2022])and Pre-Trained Backbone (Whisper-LCNN [Kawa *et al.*, 2023], Wav2vec2-AASIST [Lu *et al.*, 2024a], MiO [Phukan *et al.*, 2024]) approaches. These are traditional classifier-based baselines. MiO performs the best showing the effectiveness of dual-encoder fusion due to the emergence of complementary behavior amongst them. Then we present the results of finetuning large ALMs, we fine-tune Qwen2-Audio-Base and SeaLLMs-Audio-7B with LoRA with the same configuration as used for **GARUDA** for fair comparison. We consider Qwen2-Audio-Base and SeaLLMs-Audio-7B for finetuning experiments as they have shown relatively better performance than other large ALMs in zero-shot setting. We see that in-domain training on mixture of SEA-CF and CodecFake [Lu *et al.*, 2024a] generally improves performance over their zero-shot performances as well as traditional classifier-based baselines. These shows the importance of in-domain training as well as effectiveness of forming the deepfake detection task as a audio question answering task.

In-domain Training: GARUDA: We first report results under training setting (i), where only the projection modules are trained. As shown in Table 1 under the **GARUDA** section, **GARUDA** consistently outperforms fine-tuned Qwen2-Audio-Base and SeaLLMs-Audio-7B, despite being significantly more lightweight and without any decoder fine-tuning. Notably, both Qwen2-Audio-Base and SeaLLMs-Audio-7B are built upon 7B-parameter LLM backbones. These results highlight the effectiveness of dual-encoder fusion driven by complementary audio representations and the proposed JS alignment loss. Furthermore, our findings reinforce prior observations that audio encoders constitute the primary performance bottleneck in ALMs, while the choice of the LM decoder plays a comparatively less critical role than the effective selection and fusion of audio encoders [Li *et al.*, 2025]. We additionally conduct ablation studies by selectively removing components of **GARUDA** to assess their individual contributions. Only Wh and Only XV correspond to configurations where a single audio encoder (Whisper or x-vector) is used within the **GARUDA** framework while keeping the remaining modeling unchanged. Wh+XV (Concat), Wh+XV (CA), Wh+XV (KL) denote fusion via simple concatenation without JS alignment loss, cross-attention and KL divergence alignment, respectively. While these variants outperform traditional classifier-based baselines, their performance remains inferior to the fine-tuned Qwen2-Audio-Base. This observation underscores the importance of effective audio encoder fusion for achieving optimal performance as shown by **GARUDA**. Secondly, we report results under training setting (ii), where the LM decoder is fine-tuned. The corresponding results are summarized in Table 1 under the **GARUDA-FT** section. Compared to training setting (i), we observe a consistent performance improvement, as fine-tuning the decoder mitigates hallucination effects and enables better alignment of the model with the downstream task.

Training on Segments of Data: We further conduct experiments by training the models on subsets of the training data, following the protocol adopted in [Gu *et al.*, 2025]. The results are reported in Table 1 under the Training on Segments of Data section. We evaluate Qwen2-Audio-Base-FT (fine-tuned Qwen2-Audio-Base), **GARUDA** under training setting (i), and **GARUDA-FT** under training setting (ii). Specifically, models are trained using 75%, 50%, and 25% of the training data and evaluated on the full test sets of SEA-CF and CodecFake [Lu *et al.*, 2024a]. Across all data regimes, **GARUDA-FT** consistently achieves the best performance, demonstrating strong robustness under limited-data training conditions.

Evaluation on Unseen Test Set: All results discussed above correspond to evaluation setting (i), where the test data is drawn from seen conditions. We further evaluate our models under evaluation setting (ii), which considers unseen test conditions. The corresponding results are reported in Table 2. Using models trained on the full training set, **GARUDA** remains consistently strong under held-out conditions, outperforming both traditional classifier-based baselines and the ALM baseline.

Summary: Overall, our experimental results demonstrate that fusing complementary audio encoders and fine-tuning the LM decoder leads to superior performance compared to fine-tuning the decoder with suboptimal audio encoders (e.g., Qwen2-Audio-Base and SeaLLMs-Audio-7B) or training without decoder fine-tuning.

Lightweight and Efficiency Analysis: We perform an efficiency analysis of **GARUDA**. **GARUDA** consists of Qwen2-0.5B LM decoder. Their audio encoders, Projection and LoRA parameters sums up to be approx 110M so, the total parameters still less than 1B far less then Qwen2-Audio-Base and SeaLLMs-Audio-7B. Also, to quantify, JS adds no parameters on top of simple concatenation based fusion as it is a loss func-

tion and no trainable parameters. Further, we give an latency analysis over the whole seen training (test) set, Qwen2-Audio-Base (Finetuned) incurs an average inference time of 12.32 seconds whereas **GARUDA-FT** only 1.21s. To sum up, **GARUDA** is effective framework for CF detection in SEA-CF and beyond while being extremely efficient in both model size as well as inference time.

Comparison to previous SOTA models on the CodecFake benchmark [Lu *et al.*, 2024a]: Since AASIST, Wav2vec2-AASIST, and MiO represent prior SOTA approaches for speech deepfake detection, including CF detection, our small-ALM framework, **GARUDA**, establishes a new SOTA on the CodecFake dataset [Lu *et al.*, 2024a].

Statistical significance Test: We evaluate the statistical reliability of the observed performance gains using a paired McNemar's test on the same test sets employed in prior audio deepfake detection studies [Batra *et al.*, 2025]. Specifically, we compare **GARUDA-FT** against the strongest baselines, MiO and Qwen2-Audio-Base (Finetuned), using paired predictions on identical evaluation samples. **GARUDA-FT** shows statistically significant improvements over both MiO and Qwen2-Audio-Base (Finetuned) across SEA-CF and CodecFake [Lu *et al.*, 2024a] benchmarks. The reported p-values correspond to pairwise McNemar tests comparing **GARUDA-FT** against each baseline: 0.0047 (vs MiO) and 0.00021 (vs Qwen2-Audio-Base (Finetuned)) on SEA-CF, and 0.0003 (vs MiO) and 0.00019 (vs Qwen2-Audio-Base (Finetuned)) on CodecFake [Lu *et al.*, 2024a], all well below the 0.05 threshold. This indicates that the observed performance gains of **GARUDA-FT** are consistent and not due to random variation.

6 Conclusion

In this work, we introduce SEA-CF, the first large-scale benchmark specifically designed for CF detection in SEA languages. SEA-CF covers multiple languages, diverse speaker profiles, and a broad spectrum of SOTA NAC architectures. Our experiments reveal that existing SOTA CF detectors trained predominantly on English-centric datasets exhibit poor generalization to SEA speech, underscoring the necessity of in-domain training for SEA languages. To address this gap, we propose **GARUDA**, a novel small-ALM that achieves SOTA performance compared to both traditional classifier-based CF detectors and large ALM baselines such as Qwen2-Audio-Base and SeaLLMs-Audio-7B. Despite being significantly more lightweight, **GARUDA** consistently delivers superior performance while requiring substantially fewer parameters and lower inference latency. Overall, our work establishes a strong foundation for future research on CF detection in low-resource SEA languages and beyond. Furthermore, our work also presents an efficient and practical framework for real-world applicability.

Limitations and Future Work: Despite being the first effort toward constructing a CF benchmark for SEA languages, our work has several limitations. First, SEA-CF does not yet cover all SEA languages; we plan to extend the benchmark to additional languages in future work. Second, as the SEA-Spoof dataset has not yet been released, our evaluation is currently restricted to the available benchmarks. Once SEA-Spoof becomes publicly available, we aim to develop more generalized models capable of detecting speech deepfakes across diverse generators. Future work will also explore reasoning-centric ALMs for CF detection, with a focus on equipping detectors with natural language explanation capabilities that can articulate why a given speech sample is classified as real or fake, thereby facilitating easier interpretation and adoption by end users.

7 Disclosure for Use of AI Assistants

AI assistants were used solely for language refinement. All technical contributions, experimental design, modeling, dataset construction, and analysis were carried out by the authors.

Contribution Statement

Orchid Chetia Phukan, Girish, and Mohd Mujtaba Akhtar contributed equally as first co-authors.

References

[Akhtar *et al.*, 2026a] Mohd Mujtaba Akhtar, Girish, Farhan Sheth, and Muskaan Singh. Bridging attribution and open-set detection using graph-augmented instance learning in synthetic speech. In *Proc. EACL*, pages 5404–5413, 2026.

[Akhtar *et al.*, 2026b] Mohd Mujtaba Akhtar, Girish, and Muskaan Singh. Hcfd: A benchmark for audio deepfake detection in healthcare, 2026.

[Ardila *et al.*, 2020] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, pages 4218–4222, 2020.

[Ba *et al.*, 2023] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, pages 2033–2044, 2023.

[Batra *et al.*, 2025] Arnesh Batra, Dev Sharma, Krish Thukral, Ruhani Bhatia, Naman Batra, and Aditya Gautam. Melody or machine: Detecting synthetic music with dual-stream contrastive learning. *TMLR*, 2025.

[Borsos *et al.*, 2023] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023.

[Chintha *et al.*, 2020] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE J. Sel. Top. Signal Process.*, 14(5):1024–1037, 2020.

[Chuchra *et al.*, 2026] Akanksha Chuchra, Shukesh Reddy, Sudeepta Mishra, Abhijit Das, and Abhinav Dhall. Investigating the viability of employing multi-modal large

language models in the context of audio deepfake detection. *arXiv preprint arXiv:2601.00777*, 2026.

[Cui *et al.*, 2025] Jianqiao Cui, Bingyao Yu, Qihao Wang, Fei Meng, and Jiwen Lu. Whiadd: Semantic-acoustic fusion for robust audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11610–11618, 2025.

[Défossez *et al.*, 2022] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[Défossez *et al.*, 2024] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv:2410.00037*, 2024.

[Du *et al.*, 2024] Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *Proc. ICASSP*, pages 591–595, 2024.

[Girish *et al.*, 2025] Girish, Mohd Mujtaba Akhtar, Farhan Sheth, and Muskaan Singh. Towards attribution of generators and emotional manipulation in cross-lingual synthetic speech using geometric learning. In *Proc. IJCNLP-AACL*, pages 635–645, 2025.

[Girish *et al.*, 2026] Girish, Mohd Mujtaba Akhtar, Orchid Chetia Phukan, and Arun Balaji Buduru. Indic-codecfake meets satyam: Towards detecting neural audio codec synthesized speech deepfakes in indic languages, 2026.

[Gomez-Alanis *et al.*, 2019] Alejandro Gomez-Alanis, Antonio M Peinado, Jose A Gonzalez, and Angel M Gomez. A light convolutional gru-rnn deep feature extractor for asv spoofing detection. In *Proc. Interspeech*, volume 2019, pages 1068–1072, 2019.

[Gu *et al.*, 2025] Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proc. ACM MM*, pages 11736–11745, 2025.

[Huang *et al.*, 2025] Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In *Proc. ACL*, pages 9985–9998, 2025.

[Javed *et al.*, 2023] Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. IndicSUPERB: A speech processing universal performance benchmark for indian languages. In *Proc. AAAI*, volume 37, pages 12942–12950, 2023.

[Jung *et al.*, 2022] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. ICASSP*, pages 6367–6371, 2022.

[Kawa *et al.*, 2023] Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szymański, and Piotr Syga. Improved deepfake detection using whisper features. In *Interspeech 2023*, pages 4009–4013, 2023.

[Kumar *et al.*, 2024] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.

[Li *et al.*, 2025] Yupei Li, Li Wang, Yuxiang Wang, Lei Wang, Rizhao Cai, Jie Shi, Björn W Schuller, and Zhizheng Wu. Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components. *arXiv:2512.08403*, 2025.

[Liu *et al.*, 2023] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:2507–2522, 2023.

[Liu *et al.*, 2025] Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. Seallms-audio: Large audio-language models for southeast asia. *arXiv:2511.01670*, 2025.

[Lu *et al.*, 2024a] Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, et al. Codecfake: An initial dataset for detecting LLM-based deepfake audio. In *Proc. Interspeech*, 2024.

[Lu *et al.*, 2024b] Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. Codecfake: An Initial Dataset for Detecting LLM-based Deepfake Audio. In *Interspeech 2024*, pages 1390–1394, 2024.

[Luong and Vu, 2016] Hieu-Thi Luong and Hai-Quan Vu. A non-expert kaldi recipe for vietnamese speech recognition system. In *Proc. WLSI/OIAF4HLT*, pages 51–55, 2016.

[Müller *et al.*, 2024] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 IJCNN*, pages 1–7. IEEE, 2024.

[Nautsch *et al.*, 2021] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(2):252–265, 2021.

[Patel and Patil, 2015] Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Interspeech*, pages 2062–2066, 2015.

[Phukan *et al.*, 2024] Orchid Chetia Phukan, Gautam Kashyap, Arun Balaji Buduru, and Rajesh Sharma. Heterogeneity over homogeneity: Investigating multilingual

speech pre-trained models for detecting audio deepfake. In *Findings of NAACL*, pages 2496–2506, 2024.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, pages 28492–28518, 2023.

[Ranjan *et al.*, 2025] Rishabh Ranjan, Mayank Vatsa, and Richa Singh. Indicfake meets SAFARI-LLM: Unifying semantic and acoustic intelligence for multilingual deepfake detection. *TMLR*, 2025.

[Scardapane *et al.*, 2017] Simone Scardapane, Lucas Stoffl, Florian Röhrbein, and Aurelio Uncini. On the use of deep recurrent neural networks for detecting audio spoofing attacks. In *2017 IJCNN*, pages 3483–3490. IEEE, 2017.

[Sharma *et al.*, 2025] Divya V. Sharma, Vijval Ekbote, and Anubha Gupta. IndicSynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In *Proc. ACL*, pages 22037–22060, 2025.

[Sheth *et al.*, 2025] Farhan Sheth, Girish, Mohd Mujtaba Akhtar, and Muskaan Singh. Curved worlds, clear boundaries: Generalizing speech deepfake detection using hyperbolic and spherical geometry spaces. In *Proc. IJCNLP-AACL*, pages 1923–1932, 2025.

[Siuzdak *et al.*, 2024] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. Snac: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[Snyder *et al.*, 2018] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[Su *et al.*, 2025] Bo-Hao Su, Hui-Ying Shih, Jinchuan Tian, Jiatong Shi, Chi-Chun Lee, Carlos Busso, and Shinji Watanabe. Reasoning beyond majority vote: An explainable speechlm framework for speech emotion recognition. *arXiv:2509.24187*, 2025.

[Sutter *et al.*, 2020] Thomas Sutter, Imant Daunhawer, and Julia Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110, 2020.

[Suwanbandit *et al.*, 2023] Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *Proc. Interspeech*, volume 2, 2023.

[Tak *et al.*, 2022] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv:2202.12233*, 2022.

[Team and others, 2024] Qwen Team et al. Qwen2 technical report. *arXiv:2407.10671*, 2(3), 2024.

[Wang and Yamagishi, 2022] Xin Wang and Junichi Yamagishi. Investigating self-supervised front ends for speech spoofing countermeasures. In *Proc. Odyssey 2022*, pages 100–106, 2022.

[Wang *et al.*, 2024] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv:2408.08739*, 2024.

[Wu *et al.*, 2015] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Interspeech*, pages 2037–2041, 2015.

[Wu *et al.*, 2023] Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023*, pages 1–5. IEEE, 2023.

[Wu *et al.*, 2024] Haibin Wu, Yuan Tseng, and Hung-yi Lee. Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. In *Proc. Interspeech*, 2024.

[Wu *et al.*, 2025] Jinyang Wu, Nana Hou, Zihan Pan, Qiquan Zhang, Sailor Hardik Bhupendra, and Soumik Mondal. SEA-Spoof: Bridging the gap in multilingual audio deepfake detection for south-east asian, 2025.

[Xie *et al.*, 2025] Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, et al. The codecfake dataset and countermeasures for the universal detection of deepfake audio. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2025.

[Yang *et al.*, 2025] Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, et al. GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. In *Proc. ACL*, pages 2673–2686, 2025.

[Zeghidour *et al.*, 2021] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 30:495–507, 2021.

[Zhang *et al.*, 2024] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Zhu *et al.*, 2025] Yi Zhu, Heitor R Guimarães, Arthur Pimentel, and Tiago Falk. Auddt: Audio unified deepfake detection benchmark toolkit. *arXiv preprint arXiv:2509.21597*, 2025.