# IndicSynth: A Large-Scale Multilingual Synthetic Speech Dataset for Low-Resource Indian Languages

**Divya V. Sharma**
SBILab, IIIT-Delhi
divyas@iiitd.ac.in

**Vijval Ekbote**
SBILab, IIIT-Delhi
vijval22569@iiitd.ac.in

**Anubha Gupta**
SBILab, IIIT-Delhi
anubha@iiitd.ac.in

## Abstract

Recent advances in synthetic speech generation technology have facilitated the generation of high-quality synthetic (fake) speech that emulates human voices. These technologies pose a threat of misuse for identity theft and the spread of misinformation. Consequently, the misuse of such powerful technologies necessitates the development of robust and generalizable audio deepfake detection (ADD) and anti-spoofing models. However, such models are often linguistically biased. Consequently, the models trained on datasets in one language exhibit a low accuracy when evaluated on out-of-domain languages. Such biases reduce the usability of these models and highlight the urgent need for multilingual synthetic speech datasets for bias mitigation research. However, most available datasets are in English or Chinese. The dearth of multilingual synthetic datasets hinders multilingual ADD and anti-spoofing research. Furthermore, the problem intensifies in countries with rich linguistic diversity, such as India. Therefore, we introduce IndicSynth, which contains 4,000 hours of synthetic speech from 989 target speakers, including 456 females and 533 males for 12 low-resourced Indian languages. The dataset includes rich metadata covering gender details and target speaker identifiers. Experimental results demonstrate that IndicSynth is a valuable contribution to multilingual ADD and anti-spoofing research. The dataset can be accessed from https://github.com/vdivyas/IndicSynth.

## 1 Introduction

Recent advances in text-to-speech (TTS) and voice conversion (VC) models have enabled the generation of high-quality synthetic speech recordings that emulate human voices (Ba et al., 2023). Such recordings have applications in diverse domains, including assistive technologies, media, entertainment, and education (Bird and Lotfi, 2023; Rabhi et al., 2024; Yi et al., 2023). Despite these applications, there exists a severe threat of misuse of these technologies to spread fake news, malign personalities, financial fraud, or other criminal activities (Müller et al., 2024b; Ju et al., 2024). It is because these technologies can generate realistic synthetic speech recordings that can deceive both humans and security systems, such as speaker verification. Therefore, to prevent potential misuse, it is crucial to develop robust systems to detect fake (synthetic) audio generated by these technologies (Müller et al., 2024b).

An audio deepfake is a synthetic speech recording that appears natural enough to deceive humans and can potentially be used to spread misinformation that can cause public unrest and panic (Müller et al., 2024b). In contrast, audio spoofing refers to the techniques used to generate synthetic audios that mimic a target speaker's voice. Such techniques are often misused to deceive biometric security systems, such as speaker verification (Kawa et al., 2022; Müller et al., 2024b). Consequently, audio spoofing can lead to impersonation and identity theft. For instance, in 2019, imposters used spoofed audios of a corporate executive for financial fraud of 243,000 US dollars (USD) (Munir et al., 2024; Frank and Schönherr, 2021). Similarly, an audio deepfake-based cybercrime caused a loss of 35 million USD for a UAE-based company (Rabhi et al., 2024). Audio deepfakes are often misused to tarnish the reputation of celebrities and politicians and to manipulate public opinion during elections (Kawa et al., 2022). Therefore, developing robust audio deepfake detection and anti-spoofing technologies is crucial for social good.

For the development of robust audio deepfake detection (ADD) models, realistic synthetic speech datasets are vital. However, most existing datasets are in high-resource languages, such as English and Chinese (Müller et al., 2024b; Ba et al., 2023; Munir et al., 2024; Frank and Schönherr, 2021;

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22037–22060
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Khalid et al., 2021). Consequently, due to the ease of availability of datasets, most previous works on ADD focus on these languages (Yi et al., 2024). However, the ADD models trained on synthetic speech datasets in one language often exhibit a significantly lower accuracy when evaluated on out-of-domain languages (Müller et al., 2024b). Therefore, synthetic speech datasets in low-resource languages are urgently needed to enhance the global usability of these models. The need for such datasets intensifies in countries with rich linguistic diversity, such as India. India has 22 constitutionally recognized languages spoken by one billion speakers, and around 75% of Indians come across some deepfake content in a year (Javed et al., 2023; T and T, 2024).

In this work, we introduce IndicSynth, a novel large-scale multilingual synthetic speech dataset for 12 low-resourced Indian languages. IndicSynth contains approximately 4,000 hours of synthetic speech from 989 target speakers, including 456 females and 533 males. The dataset includes rich metadata covering the gender and identifiers of the target speakers. Additionally, to enhance IndicSynth's utility for multilingual audio deepfake detection (ADD) and anti-spoofing research, we partition our dataset into mimicry and diversity subsets. The mimicry subset contains synthetic audios closely mimicking the bonafide target voices. In contrast, the diversity subset contains a more diverse set of realistic synthetic voices. To generate IndicSynth, we apply state-of-the-art (SOTA) TTS and VC models on a publicly available bonafide (real) speech dataset. After generation, we investigate whether SOTA ADD models can accurately classify IndicSynth's synthetic audio as fake. Next, we evaluate the linguistic authenticity of our dataset using a SOTA language identification model. Subsequently, we evaluate whether the IndicSynth's mimicry subset can deceive SOTA speaker verification models through impersonation attacks. For the ease of reproducibility of our results, we use only publicly available models for experimentation.

Key contributions of this study are:

1. We introduce IndicSynth, a novel multilingual synthetic speech dataset for 12 low-resourced Indian languages containing approximately 4,000 hours of audio from 989 target speakers, including 456 females and 533 males. We partition the dataset into mimicry and diversity subsets.
2. We evaluate the linguistic bias in state-of-the-art audio deepfake detection (ADD) models and the vulnerability of SOTA speaker verification models to impersonation attacks from multilingual audio spoofs. Consequently, we investigate IndicSynth's utility towards generalizable ADD and anti-spoofing.
3. We qualitatively and quantitatively assess the linguistic authenticity of our dataset through t-SNE plots and a state-of-the-art language identification model.

## 2 Related Works

**Audio DeepFake Detection and Anti-Spoofing**: The proliferation of audio deepfakes and audio spoofs-related fraud prompted research communities to organize Audio DeepFake Detection (ADD) and ASVspoof challenges (Yi et al., 2022, 2023; Wang et al., 2020; Liu et al., 2023). Despite these initiatives, a critical yet underexplored challenge is the lack of generalizability of ADD and anti-spoofing models to out-of-domain scenarios (Korshunov and Marcel, 2022; Yousif et al., 2024; Xie et al., 2024; Kawa et al., 2022; Müller et al., 2024a). ADD and anti-spoofing models trained on speech datasets in one language exhibit significantly reduced accuracy when evaluated on out-of-domain languages. Such linguistic biases reduce the utility of these models (Sharma and Buduru, 2022; Sharma, 2024).

**Dearth of Datasets**: To mitigate linguistic biases in ADD and anti-spoofing, researchers have explored domain adaptation and data augmentation techniques (Ba et al., 2023; Xie et al., 2024). However, these techniques require synthetic speech datasets in the target languages. Most publicly available synthetic datasets, such as FakeAVCeleb, the ASVspoof 2019 dataset, and the ADD 2023 challenge dataset, are in English or Chinese (Yi et al., 2023; Müller et al., 2024b; Munir et al., 2024; Ba et al., 2023; Khalid et al., 2021; Wang et al., 2020). Consequently, researchers introduced the Urdu audio deepfake detection dataset that contains 16,830 spoofed audios (Munir et al., 2024). Similarly, the WaveFake dataset containing 196 hours of synthetic audios in English and Japanese was introduced (Frank and Schönherr, 2021). Additionally, the MLAAD dataset and the MLADDC datasets were introduced (Müller et al., 2024b; SHAH et al., 2024). However, these datasets do not include gender details and target speaker identifiers. Gender details are essential for studies related to gender bias in ADD (Xu et al., 2024; Ju et al., 2024; Kawa et al., 2022; Haut et al., 2022). The target speaker identifiers are crucial for developing defense mechanisms against impersonation attacks. In addition to these datasets, the MADD dataset contains 155.66 hours of synthetic audio for six languages (Qi et al., 2024). Thus, to address the dearth of publicly available large-scale multilingual synthetic speech datasets containing gender information and identifiers of the target speakers, we introduce the IndicSynth.

## 3 IndicSynth: Generation and Overview

This study introduces IndicSynth, a novel large-scale multilingual synthetic speech dataset. IndicSynth contains approximately 4,000 hours of speech recordings for 12 low-resourced target languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu, as illustrated in Figure 1. This section describes IndicSynth's generation methodology and its statistical details.

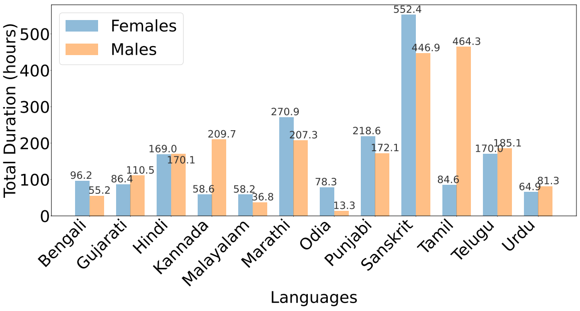

Figure 1: Total duration (in hours) of synthetic male and female voices in IndicSynth for each target language.

**IndicSynth Generation Methodology:** Firstly, we selected the IndicSuperb dataset to create the IndicSynth (Javed et al., 2023). IndicSuperb is licensed under the Creative Commons CC0 license ("no rights reserved") agreement. It contains bonafide (real) speech recordings and their associated transcripts for the 12 target languages, as shown in Figure 2. Following the bonafide dataset selection, the next step is to apply publicly available text-to-speech (TTS) based voice cloning models and voice conversion (VC) models to the bonafide IndicSuperb data[1]. TTS and VC models are widely used for synthetic data generation (Zhu et al., 2024). TTS models take a bonafide target speech recording ($v_{\text{tgt}}$) and a transcript ($t_{\text{txt}}$) as inputs. Subsequently, these models generate a synthetic speech recording (${v_{\text{tgt}}}^{\text{tts}}$) as output, as illustrated Equation 1:

$$ {v_{\text{tgt}}}^{\text{tts}} = TTS(t_{\text{txt}}, v_{\text{tgt}}) \quad (1) $$

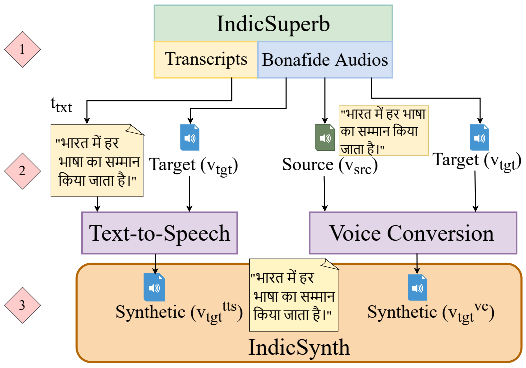

Figure 2: IndicSynth's generation methodology. We apply publicly available text-to-speech and voice conversion models to the publicly available bonafide IndicSuperb dataset to generate IndicSynth. IndicSuperb is licensed under CC0 ("no rights reserved").

The generated synthetic speech (${v_{\text{tgt}}}^{\text{tts}}$) articulates the transcript ($t_{\text{txt}}$) while attempting to mimic the target voice ($v_{\text{tgt}}$). In contrast to the TTS models, voice conversion (VC) models take a bonafide source speech recording ($v_{\text{src}}$) and a bonafide target speech recording ($v_{\text{tgt}}$) as inputs. Subsequently, these models generate a synthetic speech recording (${v_{\text{tgt}}}^{\text{vc}}$) as output, as illustrated in Equation 2:

$$ {v_{\text{tgt}}}^{\text{vc}} = VC(v_{\text{src}}, v_{\text{tgt}}) \quad (2) $$

The generated synthetic speech (${v_{\text{tgt}}}^{\text{vc}}$) articulates the transcript of the source audio ($v_{\text{src}}$), while attempting to mimic the target voice ($v_{\text{tgt}}$).

**Mimicry and Diversity:** IndicSynth includes two types of synthetic data, illustrated in Table 1:

1. *Mimicry*: The mimicry subset contains synthetic audios that closely mimic target voices. This subset is valuable for assessing the vulnerability of speaker verification systems to impersonation attacks (Munir et al., 2024).
2. *Diversity*: The diversity subset includes synthetic audios with low voice similarity to target voices. Consequently, this subset has more diversity in synthetic voices. The diversity subset is valuable because training audio deepfake detection models on such diverse, mul-

[1] Models used for IndicSynth generation: https://github.com/coqui-ai/TTS. The training datasets of these models are not fully disclosed.

| Language | Model | Category | #Females | #Male | #Clips | Duration (hrs) |
|---|---|---|---|---|---|---|
| Bengali | XTTS-v2 | Mimicry | 18 | 10 | 28,056 | 50.67 |
| | FreeVC24 | Diversity | 18 | 10 | 27,336 | 51.46 |
| | VITS | Diversity | 18 | 10 | 28,056 | 49.30 |
| Gujarati | XTTS-v2 | Mimicry | 25 | 34 | 59,118 | 97.22 |
| | FreeVC24 | Diversity | 25 | 34 | 59,660 | 99.70 |
| Hindi | XTTS-v2 | Diversity | 53 | 48 | 101,202 | 171.36 |
| | FreeVC24 | Diversity | 53 | 48 | 104,736 | 167.77 |
| Kannada | XTTS-v2 | Mimicry | 13 | 43 | 55,611 | 127.44 |
| | FreeVC24 | Diversity | 16 | 43 | 59,412 | 140.89 |
| Malayalam | XTTS-v2 | Mimicry | 10 | 7 | 17,034 | 46.35 |
| | FreeVC24 | Diversity | 10 | 7 | 17,094 | 48.61 |
| Marathi | XTTS-v2 | Mimicry | 51 | 72 | 123,246 | 231.74 |
| | FreeVC24 | Diversity | 51 | 72 | 130,150 | 246.461 |
| Odia | XTTS-v2 | Mimicry | 22 | 4 | 26,052 | 45.34 |
| | FreeVC24 | Diversity | 22 | 4 | 26,184 | 46.30 |
| Punjabi | XTTS-v2 | Mimicry | 67 | 55 | 122,244 | 191.60 |
| | FreeVC24 | Diversity | 67 | 55 | 126,110 | 199.11 |
| Sanskrit | XTTS-v2 | Diversity | 100 | 85 | 185,370 | 422.862 |
| | FreeVC24 | Diversity | 100 | 85 | 192,134 | 576.21 |
| Tamil | XTTS-v2 | Mimicry | 32 | 106 | 138,276 | 280.42 |
| | FreeVC24 | Diversity | 32 | 106 | 144,036 | 298.42 |
| Telugu | XTTS-v2 | Mimicry | 41 | 43 | 84,168 | 175.65 |
| | FreeVC24 | Diversity | 41 | 43 | 85,728 | 179.41 |
| Urdu | XTTS-v2 | Mimicry | 21 | 26 | 47,094 | 72.34 |
| | FreeVC24 | Diversity | 21 | 26 | 47,804 | 73.95 |

Table 1: Overview of IndicSynth, including generative model name, subset type (category), number of male and female target speakers, number of audio clips, and total duration of synthetic audio (in hours) for each language.

tilingual synthetic datasets can enhance their generalizability to out-of-domain languages.

**IndicSynth Generation**: To generate synthetic data for the mimicry subset, we fine-tuned the XTTS-v2 model on IndicSuperb for each of the following 10 target languages: Bengali, Gujarati, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu[2]. Fine-tuning XTTS-v2 and generating synthetic data using the same bonafide dataset (IndicSuperb) helps ensure a high similarity between synthetic and target voices. In contrast, the diversity subset includes synthetic audios directly generated from TTS and VC models without fine-tuning on IndicSuperb. We utilized the publicly available Coqui VITS model (a TTS model trained on undisclosed Bengali data) to generate Bengali synthetic data for diversity subset (Eren and The Coqui TTS Team, 2021). Additionally, we generated synthetic data for each of the 12 target languages using the publicly available XTTS-v2 text-to-speech model and the FreeVC24 voice conversion model (Eren and The Coqui TTS Team, 2021), as illustrated in Table 1. The bonafide audios and transcripts were randomly chosen for IndicSynth creation. Also, we ensured that the genders of the randomly chosen source and target speakers were the same for high-quality voice conversion.

**IndicSynth Metadata:** IndicSynth contains separate metadata files for voice cloning done through each model for each target language[3]. For the TTS models, the metadata includes the target speaker ID, the ID of the bonafide target voice sample, the target speaker's gender, the transcript, and the ID of the generated synthetic audio clip. Similarly, for the VC models, the metadata includes the source speaker ID, bonafide IndicSuperb source audio clip ID, target speaker ID, bonafide IndicSuperb target audio clip ID, the gender of the speakers, and the ID of the synthetic audio clip. Such metadata is also valuable for studying gender bias in multilingual audio deepfake detection (ADD) (Xu et al., 2024; Ju et al., 2024). Overall, utilizing the bonafide IndicSuperb with the synthetic IndicSynth can facilitate multilingual ADD and anti-spoofing research.

## 4 Evaluation of IndicSynth

### 4.1 IndicSynth for Audio DeepFake Detection

Audio deepfake detection (ADD) models accept a speech recording as input and determine if it is bonafide (real) or synthetic (fake). These models are vital to prevent the spread of fake news or misinformation from synthetic audio. Thus, we urgently need ADD models that are inclusive and general-

[2] Code for fine-tuning XTTS-v2: https://github.com/anhnh2002/XTTSv2-Finetuning-for-New-Languages

[3] IndicSynth's directory structure is in Appendix (A).

izable to unseen languages. A crucial first step in developing generalizable ADD models is benchmarking state-of-the-art (SOTA) models on unseen language datasets. Benchmarking can help investigate potential biases in these models. Therefore, we used IndicSynth to benchmark three state-of-the-art (SOTA) publicly available ADD models: Aasist, Aasist-L, and RawNet2 (Jung et al., 2022; Tak et al., 2021)[4]. The benchmark ADD models are trained on an English dataset (LA partition of the ASVspoof 2019 challenge) (Wang et al., 2020). We evaluated them on IndicSynth without fine-tuning.

**Setup**: We created separate test sets for each target language and generative model, as illustrated in Table 2. Each set includes randomly chosen 4000 bonafide female voice samples, 4000 bonafide male voice samples, 4000 synthetic female voice samples, and 4000 synthetic male voice samples. The bonafide data was taken from IndicSuperb, whereas synthetic data was taken from IndicSynth. We refer to them as IndicSynth-IndicSuperb sets.

**Evaluation Metric:** False Match Rate (FMR) and False Non-Match Rate (FNMR) are widely used metrics for evaluating biometric systems. FMR is the rate at which an ADD model incorrectly classifies synthetic audios as bonafide. In contrast, FNMR is the rate at which an ADD model incorrectly classifies bonafide audios as synthetic. The FMR and FNMR values of an ADD model vary with classification thresholds. At a particular classification threshold, FMR becomes equal to FNMR. The value of FMR when it becomes equal to the FNMR is known as the Equal Error Rate (EER). Equal Error Rate (EER) is a standard evaluation metric for audio deepfake detection systems (Wang et al., 2020; Liu et al., 2023). Therefore, we benchmarked ADD models using EER(%). Lower EER indicates that the models accurately distinguished between bonafide and synthetic audios.

**Observations:** The Aasist and Aasist-L achieve an EER of 0.83% and 0.99% on the LA evaluation set of the ASVspoof 2019 (Jung et al., 2022). Similarly, the RawNet-2 achieved an EER of 22.38% in the DF track of the ASVspoof 2021 (Liu et al., 2023). However, as illustrated in Table 2, these benchmark models achieved significantly higher EERs on IndicSynth-IndicSuperb test sets. Elevated EERs on these test sets indicate that the models struggled to distinguish between bonafide and synthetic clips. Furthermore, we plotted the Receiver Operating Characteristic (ROC) curves as illustrated in Figure 3. The ROC curve demonstrates that the ADD models achieve extremely low Area Under the Curve (AUC) scores on the Malayalam test set created using the synthetic clips obtained from the XTTS-v2 model. Low AUC indicates poor discriminative power of the models. These ob-

| | | **EER (%)** | | |
|---|---|---|---|---|
| **Language** | **G. Model** | **Aasist** | **Aasist-L** | **RawNet-2** |
| Bengali | XTTS-v2 | 70.125 | 56.150 | **56.737** |
| | FreeVC24 | 87.963 | 86.563 | 53.537 |
| | VITS | **93.200** | **89.363** | 48.287 |
| Gujarati | XTTS-v2 | 65.050 | 55.150 | 50.113 |
| | FreeVC24 | **86.163** | **86.888** | **53.425** |
| Hindi | XTTS-v2 | 42.013 | 45.438 | 14.525 |
| | FreeVC24 | **81.775** | **81.913** | **48.513** |
| Kannada | XTTS-v2 | 55.563 | 49.188 | 42.425 |
| | FreeVC24 | **73.30** | **78.950** | **50.874** |
| Malayalam | XTTS-v2 | 67.575 | 55.013 | 46.888 |
| | FreeVC24 | **85.825** | **83.600** | **55.675** |
| Marathi | XTTS-v2 | 56.712 | 52.825 | 48.037 |
| | FreeVC24 | **79.512** | **81.587** | **52.525** |
| Odia | XTTS-v2 | 57.488 | 51.575 | **48.487** |
| | FreeVC24 | **78.888** | **82.975** | 44.350 |
| Punjabi | XTTS-v2 | 57.575 | 52.863 | 47.225 |
| | FreeVC24 | **81.925** | **82.225** | **53.775** |
| Sanskrit | XTTS-v2 | 33.438 | 38.775 | 7.95 |
| | FreeVC24 | **84.238** | **86.563** | **58.35** |
| Tamil | XTTS-v2 | 61.725 | 51.188 | 51.650 |
| | FreeVC24 | **81.700** | **83.138** | **54.999** |
| Telugu | XTTS-v2 | 54.275 | 52.700 | 46.275 |
| | FreeVC24 | **75.650** | **79.000** | **52.787** |
| Urdu | XTTS-v2 | 62.763 | 55.363 | 49.250 |
| | FreeVC24 | **78.088** | **79.825** | **50.438** |

Table 2: Benchmarking audio deepfake detection (ADD) models in IndicSynth-IndicSuperb test sets without domain adaptation. For a given target language and a particular ADD model, the highest Equal Error Rate (EER%) achieved across various generative models is highlighted in bold. When evaluated without domain adaptation, the benchmark ADD models achieve elevated EER% on IndicSynth-IndicSuperb test sets. Training ADD models on multilingual synthetic datasets, such as IndicSynth, can enhance their generalizability.

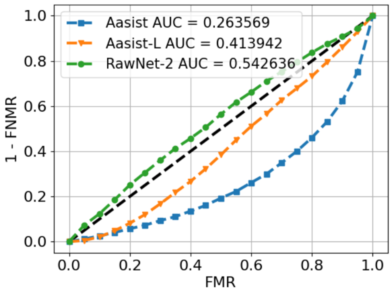

Figure 3: Receiver Operating Characteristic (ROC) Curve for Malayalam IndicSynth-IndicSuperb test set created using XTTS-v2. Low Area Under the Curve (AUC%) indicates poor discriminative power of ADD models.

[4] Aasist: `https://github.com/clovaai/aasist`
RawNet2: `https://github.com/asvspoof-challenge/2021/tree/main/DF/Baseline-RawNet2`

servations demonstrate a significant performance degradation of benchmark ADD models on unseen language test sets[5]. Training such models on large-scale multilingual synthetic speech datasets can potentially enhance their generalizability (Müller et al., 2024b). Therefore, IndicSynth is a valuable contribution towards generalizable ADD.

### 4.2 Linguistic Authenticity of IndicSynth

Next, we investigate whether IndicSynth's synthetic speech recordings accurately capture the linguistic traits of the target languages. For this experiment, we created bonafide (IndicSuperb) and synthetic (IndicSynth) test sets for each generative model and target language, as illustrated in Table 3. Each test set contains 8,000 audio clips. These sets contain an equal number of male and female voice samples. Subsequently, we evaluated IndicSynth's linguistic authenticity by running the publicly available VoxLingua107 ECAPA-TDNN spoken language identification model (Valk and Alumäe, 2021; Ravanelli et al., 2021) on these sets[6]. The model is trained on the VoxLingua107 dataset, which includes speech recordings from 107 languages (Valk and Alumae, 2021).

**Observations**: Table 3 illustrates the accuracy achieved on running the language identification model through the test sets. We observe an accuracy of more than 80% for most sets. Additionally, we compared the accuracy difference between bonafide and synthetic test sets defined as: $\Delta\ \text{Accuracy}\%=\text{Accuracy}_{\text{synthetic}}-\text{Accuracy}_{\text{bonafide}}$. For most languages, the accuracy drop is below 10%. Interestingly, the accuracies of Bengali synthetic audios from FreeVC24 and VITS are higher than bonafide audios, which indicates that these models are trained on diverse Bengali datasets.

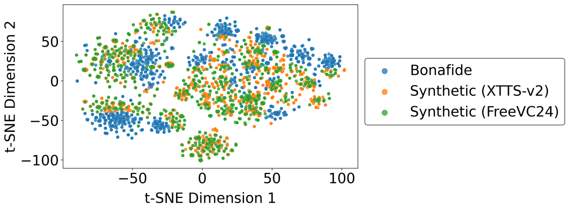

Figure 4: t-SNE visualization of bonafide (IndicSuperb) and synthetic (IndicSynth) Odia dataset. The plot indicates that the IndicSynth-Odia subset has effectively captured the linguistic traits of Odia.

| Language | Source | Accuracy | $\Delta$ Accuracy (%) |
|---|---|---|---|
| Bengali | Bonafide | 89.925 | - |
| | XTTS-v2 | 89.763 | −0.162 |
| | FreeVC24 | 90.338 | **+0.413** |
| | VITS | 98.425 | **+8.500** |
| Gujarati | Bonafide | 98.612 | - |
| | XTTS-v2 | 96.762 | −1.850 |
| | FreeVC24 | 96.475 | −2.137 |
| Hindi | Bonafide | 92.250 | - |
| | XTTS-v2 | 86.175 | −6.075 |
| | FreeVC24 | 85.525 | −6.725 |
| Kannada | Bonafide | 88.550 | - |
| | XTTS-v2 | 84.800 | −3.750 |
| | FreeVC24 | 85.638 | −2.912 |
| Malayalam | Bonafide | 97.425 | - |
| | XTTS-v2 | 96.200 | −1.225 |
| | FreeVC24 | 96.362 | −1.063 |
| Marathi | Bonafide | 94.900 | - |
| | XTTS-v2 | 89.725 | −5.175 |
| | FreeVC24 | 89.950 | −4.950 |
| Punjabi | Bonafide | 78.388 | - |
| | XTTS-v2 | 66.600 | −11.788 |
| | FreeVC24 | 65.938 | −12.450 |
| Sanskrit | Bonafide | 41.350 | - |
| | XTTS-v2 | 9.050 | −32.300 |
| | FreeVC24 | 9.175 | −32.175 |
| Tamil | Bonafide | 97.500 | - |
| | XTTS-v2 | 94.500 | −3.000 |
| | FreeVC24 | 94.812 | −2.688 |
| Telugu | Bonafide | 98.625 | - |
| | XTTS-v2 | 96.100 | −2.525 |
| | FreeVC24 | 95.638 | −2.987 |
| Urdu | Bonafide | 39.900 | - |
| | XTTS-v2 | 33.975 | −5.925 |
| | FreeVC24 | 33.763 | −6.137 |

Table 3: Language identification results. We evaluated IndicSynth's linguistic authenticity by running language identification model through various test sets for each generative model and target language (except Odia). We observe above 80% accuracy in most test sets.

**Qualitative evaluation**: The VoxLingua107 ECAPA-TDNN spoken language identification model does not support Odia. However, the training set of the model covers 107 languages. Therefore, its embeddings should capture linguistic traits effectively. Thus, we obtained the 256-dimensional language identification embeddings of the Odia test sets and visualized them through t-SNE using a perplexity of 40 (Munir et al., 2024), as shown in Figure 4. Figure 4 shows no clear separation between the bonafide (IndicSuperb) and the synthetic (IndicSynth) embeddings. The plot indicates that IndicSynth's Odia subset has effectively captured linguistic traits of Odia[7].

### 4.3 Utility of the Mimicry Subset

Speaker verification systems accept two speech recordings as input and determine if they are from the same speaker. The input speech recordings

[5] Appendix (B) includes additional plots.

[6] Language identification model: https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa

[7] The t-SNE plots for Punjabi, Sanskrit, and Urdu are in Appendix (D)

| Language | SV Model | Female Test Set | Male Test Set | Combined Test Set |
|---|---|---|---|---|
| Bengali | ECAPA-TDNN | **31.580** | 29.180 | 31.110 |
| | ResNet TDNN | **23.960** | 22.960 | 25.160 |
| | X-Vector | 43.860 | **44.520** | 43.639 |
| Gujarati | ECAPA-TDNN | 23.280 | **34.440** | 28.880 |
| | ResNet TDNN | 18.520 | **30.560** | 24.730 |
| | X-Vector | 31.040 | **32.640** | 32.200 |
| Kannada | ECAPA-TDNN | **25.880** | 22.700 | 24.360 |
| | ResNet TDNN | **22.740** | 20.060 | 21.470 |
| | X-Vector | **35.020** | 28.740 | 31.770 |
| Malayalam | ECAPA-TDNN | 30.140 | **33.680** | 32.070 |
| | ResNet TDNN | 30.700 | **31.480** | 31.170 |
| | X-Vector | **38.460** | 38.240 | 38.520 |
| Marathi | ECAPA-TDNN | 20.620 | **28.820** | 24.710 |
| | ResNet TDNN | 17.680 | **25.140** | 21.600 |
| | X-Vector | 28.260 | **31.160** | 31.080 |
| Odia | ECAPA-TDNN | 29.300 | **36.800** | 32.940 |
| | ResNet TDNN | **23.580** | 21.420 | 25.680 |
| | X-Vector | 33.440 | **48.100** | 42.450 |
| Punjabi | ECAPA-TDNN | 25.800 | **33.420** | 29.610 |
| | ResNet TDNN | 23.360 | **30.640** | 27.050 |
| | X-Vector | 32.330 | **36.160** | 34.350 |
| Tamil | ECAPA-TDNN | 20.160 | **27.760** | 23.840 |
| | ResNet TDNN | 17.820 | **25.500** | 21.540 |
| | X-Vector | 28.280 | **31.860** | 30.070 |
| Telugu | ECAPA-TDNN | 26.380 | **27.500** | 26.930 |
| | ResNet TDNN | 23.320 | **25.140** | 24.550 |
| | X-Vector | 31.760 | **32.120** | 32.180 |

Table 4: Investigating the vulnerability of state-of-the-art (SOTA) speaker verification models (SV) against impersonation attacks. We observe elevated equal error rates (EER%) when the negative trial pairs contain IndicSynth's mimicry subset's synthetic speech recordings and the target speaker's bonafide speech from IndicSuperb. It suggests that the mimicry subset audios closely mimic the bonafide target voices. Therefore, the mimicry subset of IndicSynth is a valuable resource for enhancing the robustness of SOTA SV models.

form a trial pair. Such systems are vital in forensics, business, e-commerce, and access control mechanisms. However, the malicious use of voice cloning models may lead to the generation of synthetic speech recordings that closely mimic the target voice. Such synthetic recordings (audio spoofs) may be misused to deceive speaker verification systems, leading to impersonation attacks against the target speaker. Fine-tuning speaker verification models on multilingual synthetic speech datasets can enhance their generalizability and robustness to out-of-domain audio spoofs. Therefore, this experiment explores the utility of IndicSynth's mimicry subset for enhancing the robustness of speaker verification models. We evaluate whether the synthetic audios of mimicry subset can deceive three publicly available state-of-the-art (SOTA) speaker verification models: ECAPA-TDNN, X-Vector, and ResNet TDNN (Desplanques et al., 2020; Snyder et al., 2018; Villalba et al., 2020)[8].

**Methodology**: We created speaker verification test sets, as illustrated in Table 4. Each set contains randomly generated 20,000 trial pairs with equal positives and negatives. A positive trial pair contains two bonafide (IndicSuperb) speech recordings of the same target speaker, X. In contrast, a negative trial pair contains a bonafide (IndicSuperb) speech recording of a target speaker X and a synthetic (IndicSynth) speech recording of X. Since each set contains an equal number of male and female speaker trial pairs, we refer to them as combined test sets. Each combined test set contains 5000 bonafide female trial pairs, 5000 bonafide male trial pairs, 5000 synthetic female trial pairs, and 5000 synthetic male trial pairs. Subsequently, to evaluate gender bias in speaker verification models with respect to impersonation attacks, we also split the combined test set and created separate male and female speaker test sets.

**Evaluation Metric**: We evaluate the mimicry subset using EER. A higher EER indicates that the speaker verification model struggled to distinguish between positive and negative trial pairs. It implies that the synthetic speech recordings closely mimic the target speaker's bonafide voice sample

[8] ECAPA-TDNN:https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
X-Vector:https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
ResNet TDNN: https://huggingface.co/speechbrain/spkrec-resnet-voxceleb

in a negative trial pair.

**Observations**: The SOTA speaker verification models typically achieve an EER of less than 10% when evaluated on unseen language test sets (Akram et al., 2024; Xia et al., 2019; Mandalapu et al., 2021). However, as illustrated in Table 4, we observed significantly elevated EERs ranging from 21.470% to 43.639% on the combined test sets. Elevated EERs suggest that the speech recordings in IndicSynth's mimicry subset closely mimic the bonafide (IndicSuperb) target voices. Furthermore, we compared the EERs of the male and female speaker test sets. The EERs of male and female speaker test sets for the Bengali, Malayalam, and Telugu test sets are comparable. However, the EER values for Kannada female test sets are higher than the male test sets (with absolute differences of 2.68% to 6.28% across the speaker verification models). This observation indicates that the Kannada female voices mimic the target speakers more closely than the Kannada male voices in IndicSynth. Similarly, the EERs of the male test sets are higher than the female test sets for Gujarati, Marathi, Odia, Punjabi, and Tamil. This observation indicates that in IndicSynth, male voices in these languages closely mimic the target speakers compared to female voices.

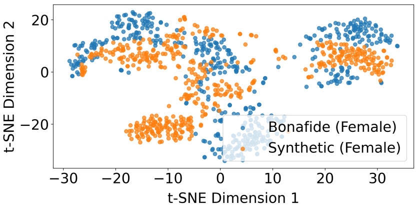

Figure 5: The t-SNE plot of bonafide (IndicSuperb) Odia and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

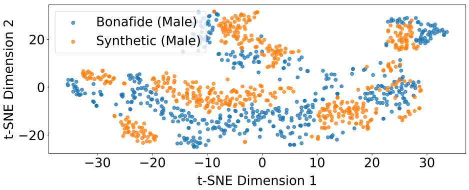

Figure 6: The t-SNE plot of bonafide (IndicSuperb) Odia and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

**Qualitative Evaluation:** For an extensive evaluation, we visualized the proximity of the bonafide (IndicSuperb) and IndicSynth's mimicry subset through t-SNE. Figure 5 and Figure 6 represent the t-SNE plots for Odia female and male voices. For each plot, we randomly sampled 500 bonafide and 500 synthetic clips of the same target speakers. Next, we created t-SNE plots with a perplexity of 40 using 80-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features of these audios (Munir et al., 2024). MFCCs are biologically inspired speech features that mimic the human auditory system. The proximity of the bonafide and synthetic embeddings in t-SNE indicates that the mimicry subset's synthetic audios closely mimic the bonafide target voices[9].

## 5 Discussion

This section reflects on our rationale behind creating mimicry and diversity subsets in IndicSynth. Additionally, we briefly review the potential utility of these subsets with our experimental results.

**Rationale behind Mimicry Subset:** IndicSynth's mimicry subset consists of synthetic voices closely mimicking the target speaker's bonafide voice. Such synthetic audios—also called audio spoofs—can deceive speaker verification systems, leading to impersonation attacks. Section 4.3 demonstrates that speaker verification systems are vulnerable to multilingual audio spoofs (IndicSynth's mimicry subset). This observation underscores the potential utility of the mimicry subset for developing multilingual anti-spoofing technologies.

**Need for a Diversity Subset:** Synthetic speech is not only misused for impersonation but also for spreading misinformation. Synthetic voices circulating in social media to spread misinformation often do not mimic a specific target speaker's voice. Instead, misinformation campaigns often involve diverse synthetic voices. Therefore, training audio deepfake detection (ADD) models on a broad range of synthetic voices is crucial for enhancing the robustness of these models. However, as we know, the mimicry subset has a limited speaker diversity as it only includes voices mimicking IndicSuperb speakers. Therefore, we introduced the diversity subset in IndicSynth to incorporate a broader range of synthetic voices beyond speaker mimicry into our dataset. Table 1 provides an overview of mimicry and diversity subsets.

**Utility of Diversity Subset:** As described in Section 4.1, we evaluated three state-of-the-art (SOTA) audio deepfake detection (ADD) models on Indic-

[9] Additional t-SNE plots are in Appendix (C).

Synth (both diversity and mimicry subsets) without domain adaptation. These ADD models achieve low Equal Error Rates (EERs) on ASVSpoof challenge datasets. However, we observed a significant performance degradation of these models on IndicSynth test sets, as illustrated by elevated EERs and low Area Under the Curve (AUC) in Table 2 and Figure 3. Munir et al. (2024) also reported a similar observation. In their work, the authors proposed an Urdu audio deepfake detection dataset. Such observations indicate a lack of generalizability of existing audio deepfake detection models to out-of-domain languages. It suggests that training or fine-tuning audio deepfake detection models on multilingual ADD datasets can potentially enhance the generalizability of these models to out-of-domain languages. Thus, bonafide IndicSuperb data combined with the synthetic IndicSynth data (diversity and mimicry subsets) can potentially serve as a valuable dataset for multilingual audio deepfake detection.

## 6 Conclusions and Future Work

This paper introduces IndicSynth, a novel large-scale multilingual synthetic speech dataset to facilitate multilingual audio deepfake detection and anti-spoofing research. The dataset contains about 4,000 hours of synthetic audio from 989 target speakers, including 456 females and 533 males for 12 low-resourced Indian languages. Additionally, IndicSynth includes rich metadata covering the identifiers and gender information of the target speakers. Thus facilitating research on gender biases in audio deepfake detection and anti-spoofing. The dataset consists of mimicry and diversity subsets. The mimicry subset includes synthetic audios that closely mimic bonafide target voices. In contrast, the diversity subset contains a diverse set of realistic synthetic voices. Experimental results demonstrate that the synthetic audios of the mimicry subset can deceive state-of-the-art (SOTA) speaker verification models through impersonation attacks. Similarly, empirical results demonstrate poor performance of SOTA audio deepfake detection models on Indian language test sets. Furthermore, qualitative and quantitative evaluation using a SOTA language identification model validated the linguistic authenticity of our dataset. It turns out that IndicSynth is a valuable contribution to preventing impersonation attacks on speaker verification systems and facilitating multilingual audio deepfake detection research.

This work opens up several avenues for research on linguistic biases in audio deepfake detection and anti-spoofing. For instance, IndicSynth can be used to investigate the underexplored problem of gender biases in multilingual audio deepfake detection and for defense against impersonation attacks through multilingual spoofs. The dataset is licensed under the CC BY-NC 4.0 license.

## 7 Limitations

This work introduces IndicSynth, a novel, large-scale multilingual synthetic speech dataset to facilitate multlingual audio deepfake detection and anti-spoofing research. We acknowledge that our work has the following limitations:

**IndicSynth's Scope and Experimentation**: IndicSynth contains synthetic speech recordings for only 12 languages. Also, the mimicry subsets for Hindi and Sanskrit are absent in IndicSynth. In the future, the dataset can be extended by covering more low-resourced languages and more voice cloning models for dataset creation. Additionally, we evaluated IndicSynth using sample test sets. We believe that the results from these sets indicate the overall dataset quality.

**Absence of User Study**: Ideally, the naturalness of synthetic speech datasets should be evaluated through a user study. The user study participants must be proficient in the target languages for authentic results. However, for large-scale multilingual datasets, such as IndicSynth, recruiting participants proficient in low-resource languages is challenging. Thus, meeting our paper's objectives, we experimentally evaluated IndicSynth using state-of-the-art speaker verification models, audio deepfake detection models, and a language identification model. Additionally, we have included t-SNE plots in the appendix for a qualitative evaluation of our dataset.

We highlight that the challenge of recruiting participants proficient in low-resourced languages is not unique to IndicSynth. Instead, it is a common issue faced by researchers working towards generating multilingual datasets for social good. However, with around 7,000 global languages and rising cases of deepfake-related fraud, there is an urgent need for multilingual synthetic datasets to facilitate research on multilingual audio deepfake detection. Therefore, the absence of user studies should not hinder the release of such datasets. Instead, the community members with access to computational resources can contribute by constructing and releasing more multilingual datasets. Subsequently, the members who can connect to native speakers of those languages can contribute by conducting user studies to evaluate human perception of synthetic speech.

Despite these limitations, the dearth of multilingual synthetic speech datasets makes IndicSynth a valuable resource that can facilitate research on multilingual audio deepfake detection and anti-spoofing.

## 8 Ethical Considerations

Synthetic speech datasets are essential to advance audio deepfake detection and anti-spoofing research. However, we realize that such datasets can inadvertently contribute to the refinement of audio deepfake generation technologies by malicious users. Therefore, responsible management of these resources is crucial. Thus, we release IndicSynth under CC BY-NC 4.0, restricting commercial use of our dataset. Furthermore, IndicSynth is a synthetic speech dataset generated from the publicly available bonafide IndicSuperb dataset. IndicSuperb is licensed under the Creative Commons CC0 license (“no rights reserved”). The CC0 license allows users to freely build upon, reuse, or enhance the dataset without restriction. We strongly encourage the community to use IndicSynth for social good and advance research on multilingual audio deepfake detection and anti-spoofing.

## Acknowledgments

This work is supported by the Infosys Centre for Artificial Intelligence (CAI) at IIIT-Delhi. We also thank the SBILab at IIIT-Delhi for their helpful discussions and support.

## References

Ali Akram, Marija Stanojevic, Malikeh Ehghaghi, and Jekaterina Novikova. 2024. Zero-shot multilingual speaker verification in clinical trials. *ArXiv*, abs/2404.01981.

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

Jordan J. Bird and Ahmad Lotfi. 2023. Real-time detection of ai-generated speech for deepfake voice conversion. *ArXiv*, abs/2308.12734.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834. ISCA.

Gölge Eren and The Coqui TTS Team. 2021. Coqui TTS.

Joel Cameron Frank and Lea Schönherr. 2021. Wavefake: A data set to facilitate audio deepfake detection. *ArXiv*, abs/2111.02813.

Kurtis Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, Dillanie Sumanthiran, M. Rafayet Ali, and Ehsan Hoque. 2022. Demographic feature isolation for bias research using deepfakes. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 6890–6897, New York, NY, USA. Association for Computing Machinery.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:12942–12950.

Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. 2024. Improving Fairness in Deepfake Detection . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4643–4653, Los Alamitos, CA, USA. IEEE Computer Society.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371.

Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. pages 4023–4027.

Hasam Khalid, Shahroz Tariq, and Simon S. Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *ArXiv*, abs/2108.05080.

Pavel Korshunov and Sébastien Marcel. 2022. Improving generalization of deepfake detection with data farming and few-shot learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):386–397.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano

Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Hareesh Mandalapu, Thomas Møller Elbo, Raghavendra Ramachandra, and Christoph Busch. 2021. Cross-lingual speaker verification: Evaluation on x-vector method. In *Intelligent Technologies and Applications*, pages 215–226, Cham. Springer International Publishing.

Sheza Munir, Wassay Sajjad, Mukeet Raza, Emaan Abbas, Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2024. Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14470–14480, Bangkok, Thailand. Association for Computational Linguistics.

Nicolas Müller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Böttinger. 2024a. Harder or different? understanding generalization of audio deepfake detection. pages 2705–2709.

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

Xiaoke Qi, Hao Gu, Jiangyan Yi, Jianhua Tao, Yong Ren, Jiayi He, and Siding Zeng. 2024. Madd: A multi-lingual multi-speaker audio deepfake detection dataset. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 466–470.

Mouna Rabhi, Spiridon Bakiras, and Roberto Di Pietro. 2024. Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, 250:123941.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

ARTH JUHUL SHAH, Ravindrakumar M. Purohit, Dharmendra H. Vaghera, and Hemant Patil. 2024. MLADDC: Multi-lingual audio deepfake detection corpus. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Divya Sharma. 2024. EcoSpeak: Cost-efficient bias mitigation for partially cross-lingual speaker verification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 379–394, Mexico City, Mexico. Association for Computational Linguistics.

Divya Sharma and Arun Balaji Buduru. 2022. FAt-Net: Cost-effective approach towards mitigating the linguistic bias in speaker verification systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1247–1258, Seattle, United States. Association for Computational Linguistics.

David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. Spoken language recognition using x-vectors. In *The Speaker and Language Recognition Workshop (Odyssey 2018)*, pages 105–111.

Sidharth T and Guhan T. 2024. Deepfake technology in social media: Social and legal implications in india. *IJFMR*, 6(6).

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

Jorgen Valk and Tanel Alumae. 2021. Voxlingua107: A dataset for spoken language recognition. pages 652–658.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. 2020. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Wei Xia, Jing Huang, and John H.L. Hansen. 2019. Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5816–5820.

Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2024. Domain generalization via aggregation and separation for audio deepfake detection. *IEEE Transactions on Information Forensics and Security*, 19:344–358.

Ying Xu, Philipp Terhörst, Marius Pedersen, and Kiran Raja. 2024. Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 5(1):93–106.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Leling Xu, Jun Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. 2023. Add 2023: the second audio deepfake detection challenge. In *DADA@IJCAI*.

Jiangyan Yi, Chu Yuan Zhang, Jianhua Tao, Chenglong Wang, Xinrui Yan, Yong Ren, Hao Gu, and Junzuo Zhou. 2024. Add 2023: Towards audio deepfake detection and analysis in the wild. *Preprint*, arXiv:2408.04967.

Mohammed Yousif, Jonat John Mathew, Huzaifa Pallan, Agamjeet Singh Padda, Syed Daniyal Shah, Sara Adamski, Madhu Reddiboina, and Arjun Pankajakshan. 2024. Enhancing generalization in audio deepfake detection: A neural collapse based sampling and training approach. *Preprint*, arXiv:2404.13008.

Yi Zhu, Surya Koppisetti, Trang Tran, and Gaurav Bharaj. 2024. Slim: Style-linguistics mismatch model for generalized audio deepfake detection. *ArXiv*, abs/2407.18517.

## A IndicSynth Directory Structure

Figure 7 shows the directory structure of IndicSynth. The dataset contains a folder for each target language. Each language folder includes subfolders for generative models: XTTS_v2, FreeVC24, or the VITS. Each generative model folder has a metadata file and subfolders for anonymous target speaker IDs. Within each target speaker's folder, there are synthetic audio clips. The speaker IDs used in the IndicSynth are the same as those used in IndicSuperb.

```
IndicSynth/
|-- <language>/
|   |-- XTTS_v2/
|   |   |-- Male/
|   |   |   |-- <speaker_id>/
|   |   |   |   |-- <clip_id>.wav
|   |   |   |   `-- ...
|   |   |-- Female/
|   |   |   |-- <speaker_id>/
|   |   |   |   |-- <clip_id>.wav
|   |   |   |   `-- ...
|   |   `-- metadata.csv
|   |-- FreeVC24/
|   |   |-- Male/
|   |   |   |-- <speaker_id>/
|   |   |   |   |-- <clip_id>.wav
|   |   |   |   `-- ...
|   |   |-- Female/
|   |   |   |-- <speaker_id>/
|   |   |   |   |-- <clip_id>.wav
|   |   |   |   `-- ...
|   |   `-- metadata.csv
```

Figure 7: IndicSynth directory structure.

## B IndicSynth for Audio DeepFake Detection: Additional Plots

Figures 8–56 illustrate the Receiver Operating Characteristic (ROC) Curves and the Detection Error Trade-off (DET) Curves for IndicSynth test sets created using various generative models. The plots indicate that the benchmark audio deepfake detection models lack generalizability on out-of-domain (Indian) language test sets.

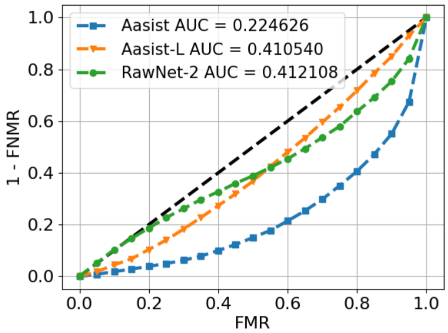

Figure 8: The Receiver Operating Characteristic (ROC) Curve for the Bengali test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

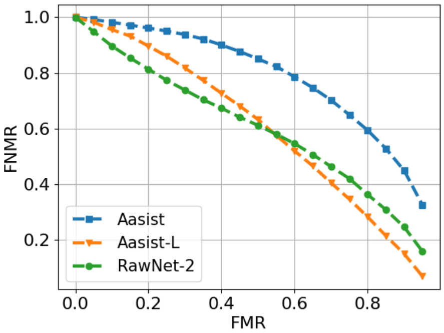

Figure 9: The Detection Error Trade-off (DET) Curve for the Bengali test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

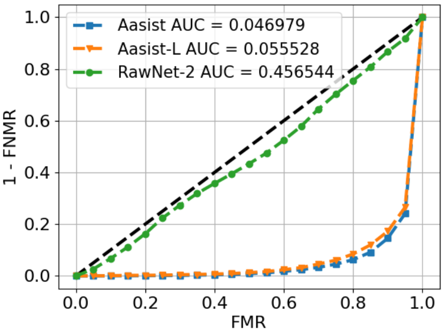

Figure 10: The Receiver Operating Characteristic (ROC) Curve for the Bengali test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

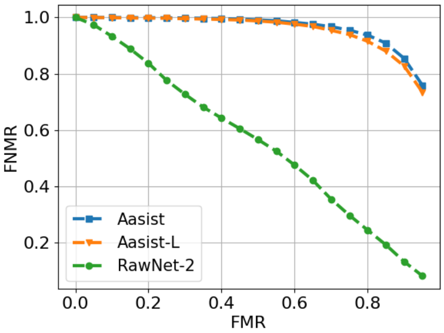

Figure 11: The Detection Error Trade-off (DET) Curve for the Bengali test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

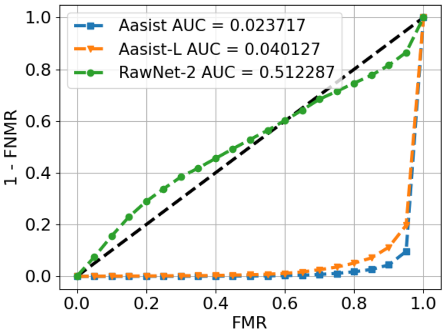

Figure 12: The Receiver Operating Characteristic (ROC) Curve for the Bengali test set created using VITS. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

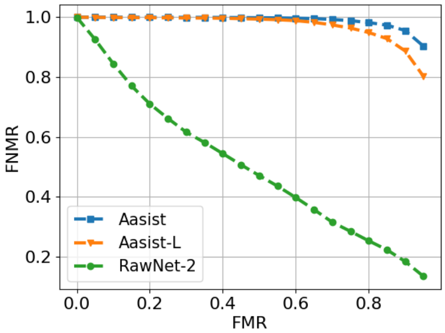

Figure 13: The Detection Error Trade-off (DET) Curve for the Bengali test set created using VITS. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

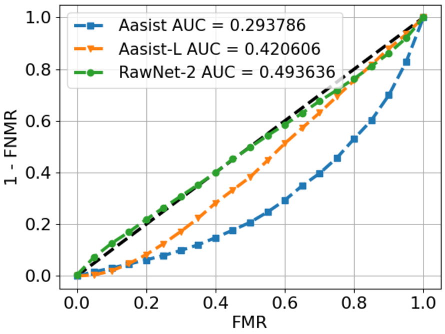

Figure 14: The Receiver Operating Characteristic (ROC) Curve for the Gujarati test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

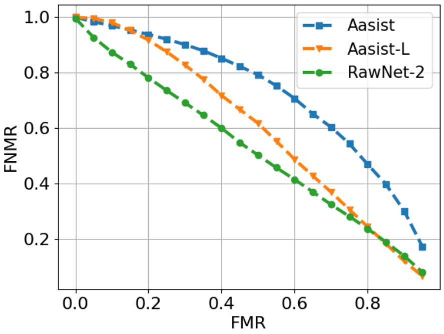

Figure 15: The Detection Error Trade-off (DET) Curve for the Gujarati test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

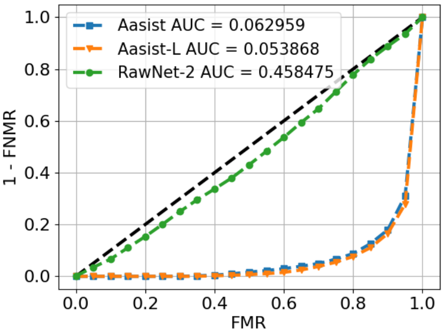

Figure 16: The Receiver Operating Characteristic (ROC) Curve for the Gujarati test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

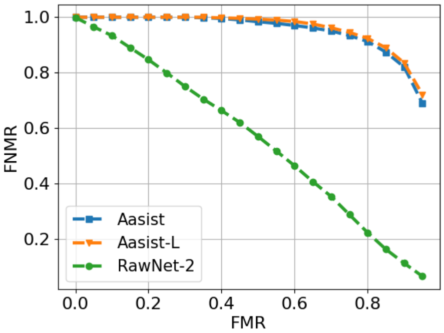

Figure 17: The Detection Error Trade-off (DET) Curve for the Gujarati test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

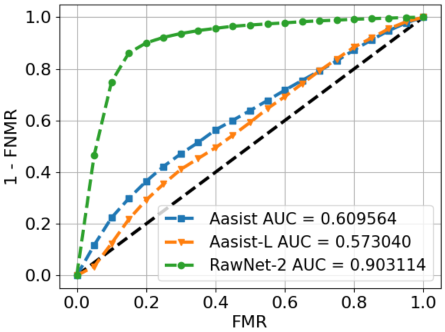

Figure 18: The Receiver Operating Characteristic (ROC) Curve for the Hindi test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

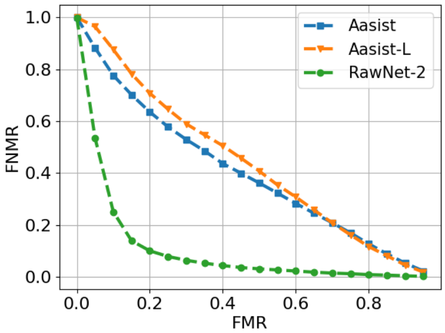

Figure 19: The Detection Error Trade-off (DET) Curve for the Hindi test set created using XTTS-v2.

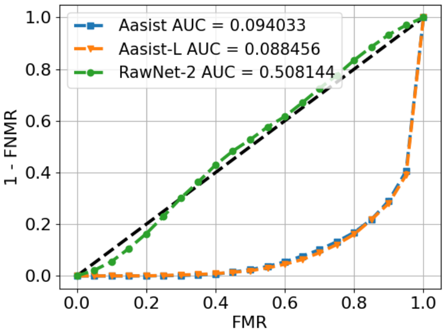

Figure 20: Receiver Operating Characteristic (ROC) Curve for the Hindi IndicSynth-IndicSuperb test set created using FreeVC24. Low Area Under the Curve (AUC%) indicates poor discriminative power of ADD models.

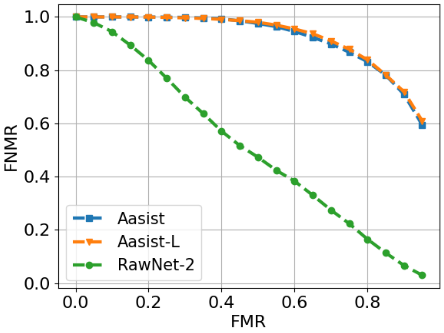

Figure 21: The Detection Error Trade-off (DET) Curve for the Hindi test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

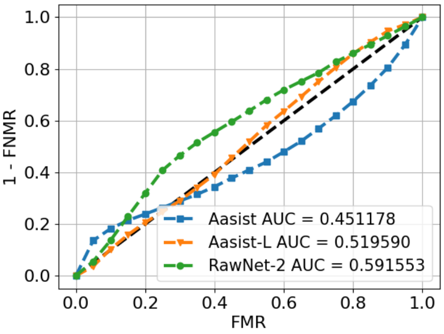

Figure 22: The Receiver Operating Characteristic (ROC) Curve for the Kannada test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

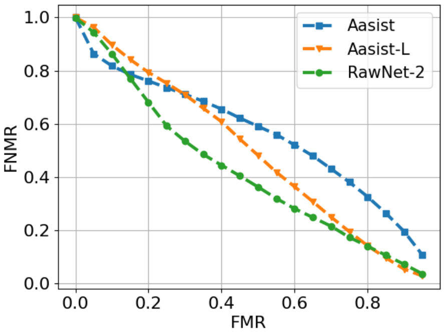

Figure 23: The Detection Error Trade-off (DET) Curve for the Kannada test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

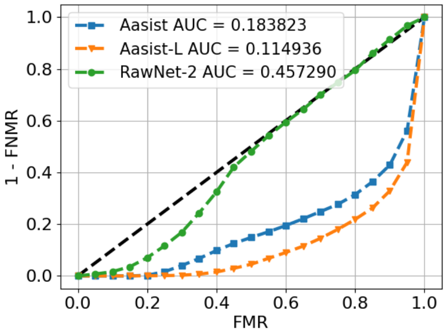

Figure 24: The Receiver Operating Characteristic (ROC) Curve for the Kannada test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

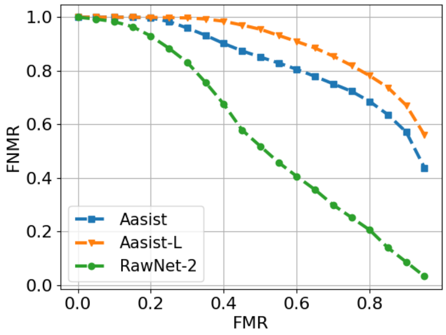

Figure 25: The Detection Error Trade-off (DET) Curve for the Kannada test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

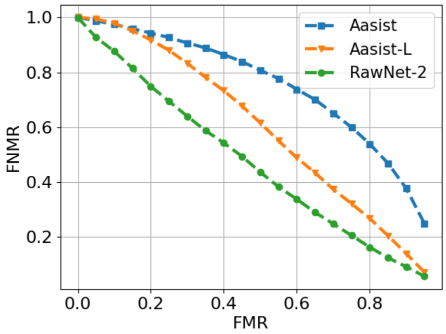

Figure 26: The Detection Error Trade-off (DET) Curve for the Malayalam test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

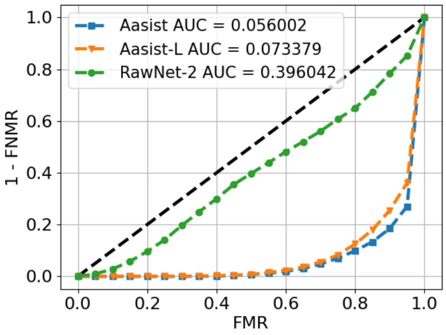

Figure 27: The Receiver Operating Characteristic (ROC) Curve for the Malayalam test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

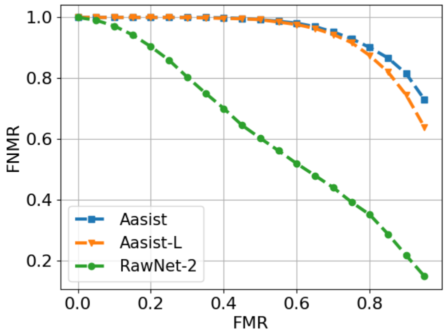

Figure 28: The Detection Error Trade-off (DET) Curve for the Malayalam test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

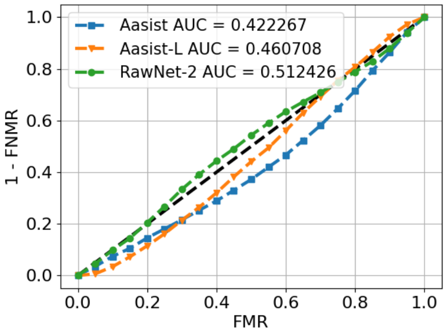

Figure 29: The Receiver Operating Characteristic (ROC) Curve for the Marathi test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

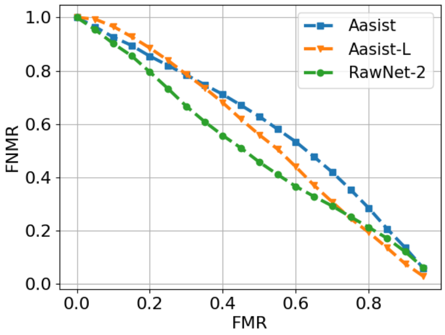

Figure 30: The Detection Error Trade-off (DET) Curve for the Marathi test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

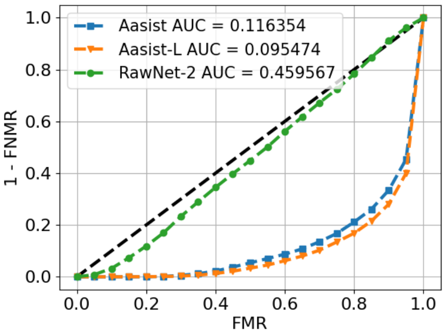

Figure 31: The Receiver Operating Characteristic (ROC) Curve for the Marathi test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

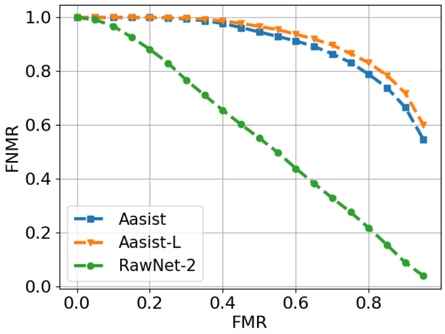

Figure 32: The Detection Error Trade-off (DET) Curve for the Marathi test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

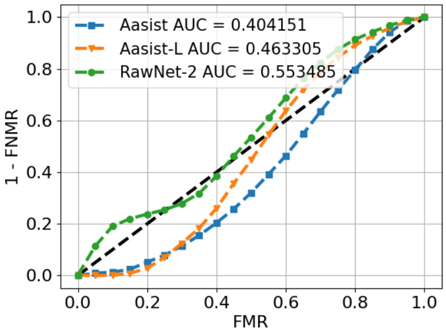

Figure 33: The Receiver Operating Characteristic (ROC) Curve for the Odia test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

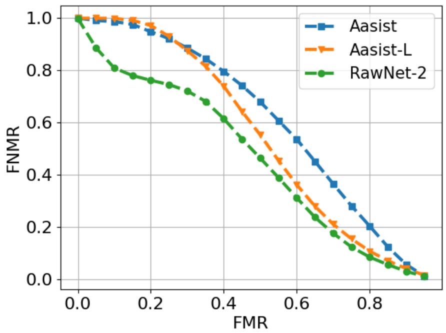

Figure 34: The Detection Error Trade-off (DET) Curve for the Odia test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

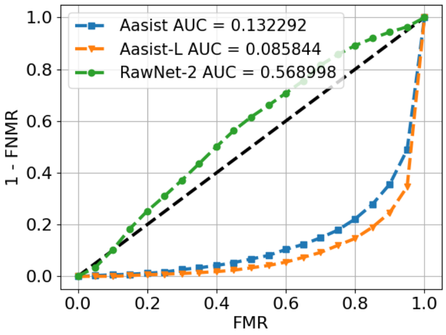

Figure 35: The Receiver Operating Characteristic (ROC) Curve for the Odia test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

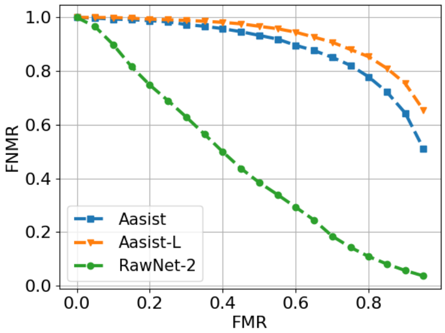

Figure 36: The Detection Error Trade-off (DET) Curve for the Odia test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

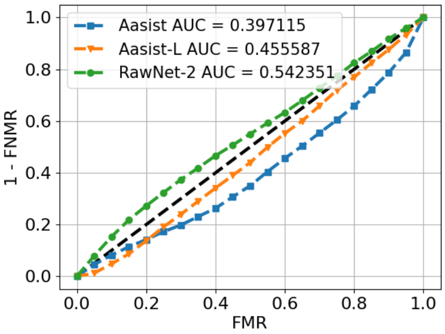

Figure 37: The Receiver Operating Characteristic (ROC) Curve for the Punjabi test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

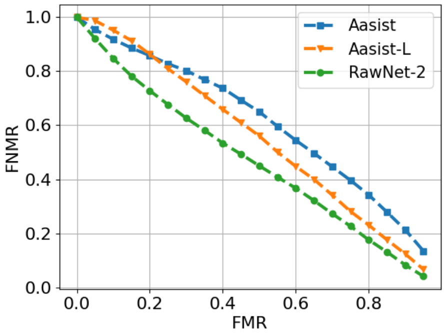

Figure 38: The Detection Error Trade-off (DET) Curve for the Punjabi test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

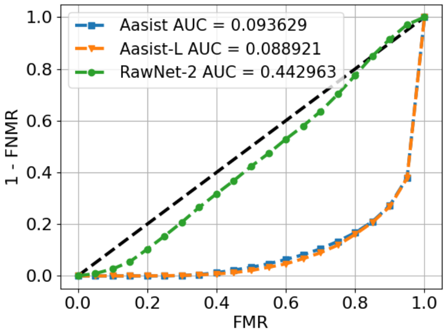

Figure 39: The Receiver Operating Characteristic (ROC) Curve for the Punjabi test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

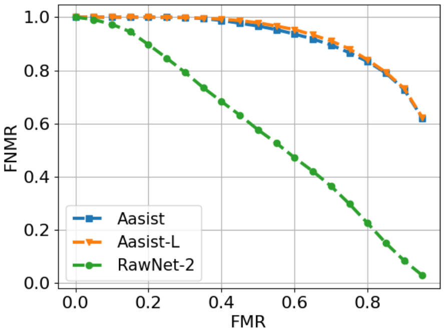

Figure 40: The Detection Error Trade-off (DET) Curve for the Punjabi test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

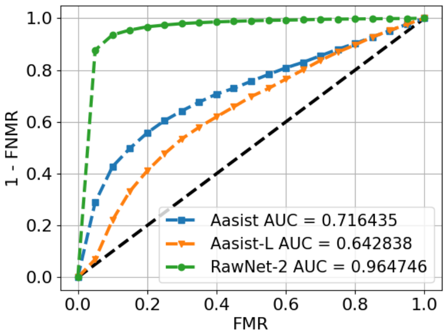

Figure 41: The Receiver Operating Characteristic (ROC) Curve for the Sanskrit test set created using XTTS-v2.

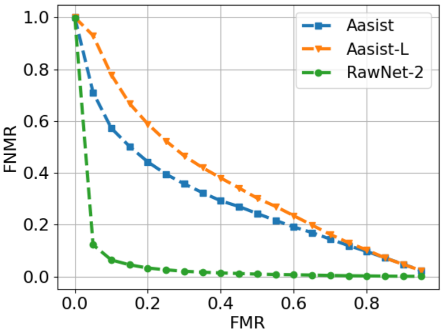

Figure 42: The Detection Error Trade-off (DET) Curve for the Sanskrit test set created using XTTS-v2. The trend of DET curves towards the bottom left indicates the capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

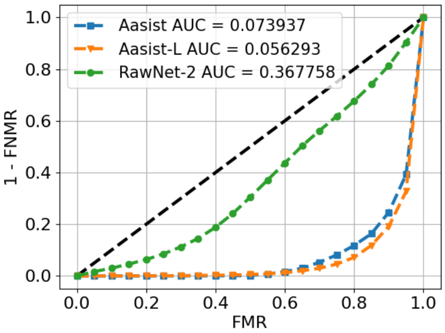

Figure 43: The Receiver Operating Characteristic (ROC) Curve for the Sanskrit test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

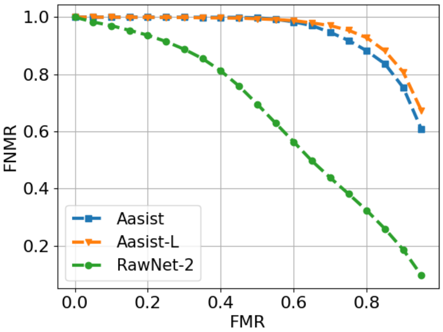

Figure 44: The Detection Error Trade-off (DET) Curve for the Sanskrit test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

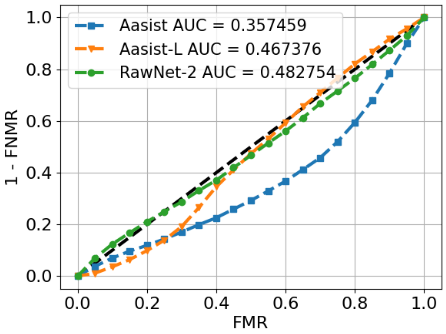

Figure 45: The Receiver Operating Characteristic (ROC) Curve for the Tamil test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

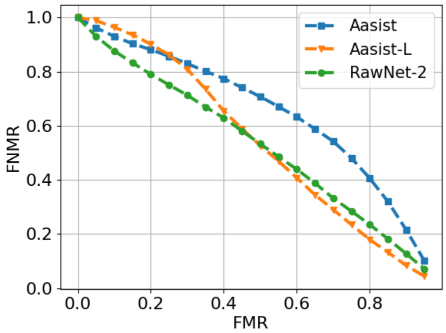

Figure 46: The Detection Error Trade-off (DET) Curve for the Tamil test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

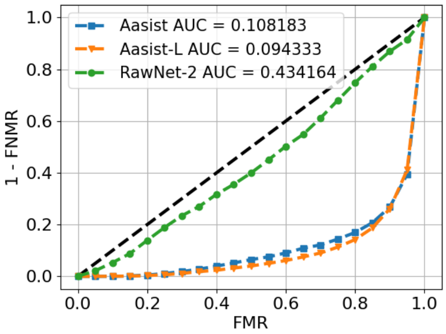

Figure 47: The Receiver Operating Characteristic (ROC) Curve for the Tamil test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

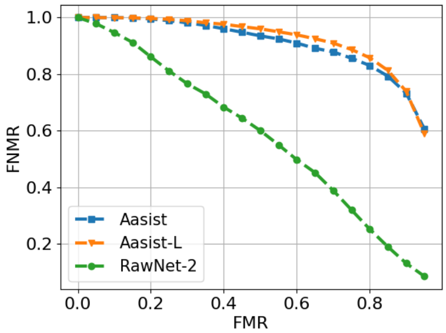

Figure 48: The Detection Error Trade-off (DET) Curve for the Tamil test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

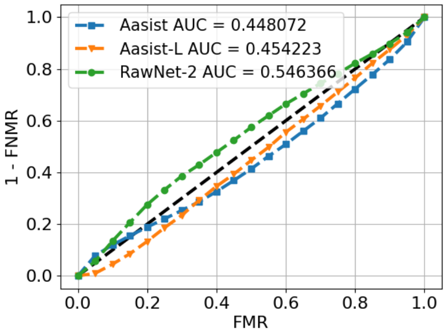

Figure 49: The Receiver Operating Characteristic (ROC) Curve for the Telugu test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

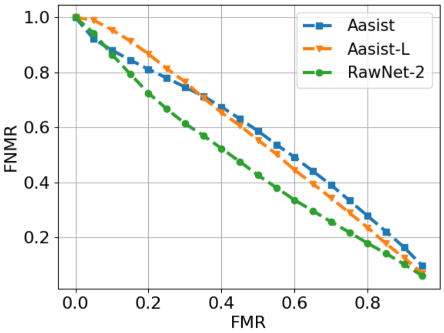

Figure 50: The Detection Error Trade-off (DET) Curve for the Telugu test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

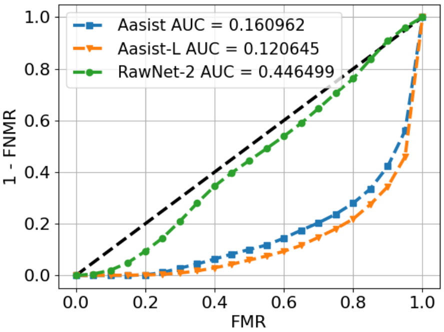

Figure 51: The Receiver Operating Characteristic (ROC) Curve for the Telugu test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

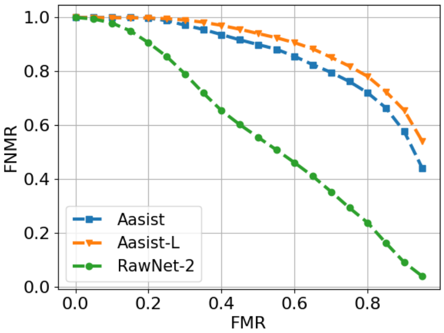

Figure 52: The Detection Error Trade-off (DET) Curve for the Telugu test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

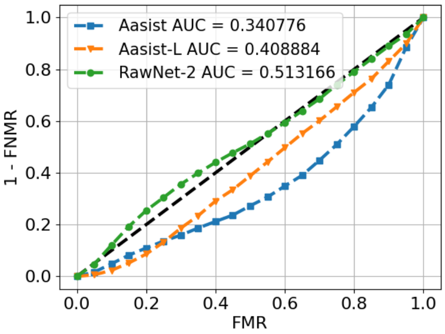

Figure 53: The Receiver Operating Characteristic (ROC) Curve for the Urdu test set created using XTTS-v2. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

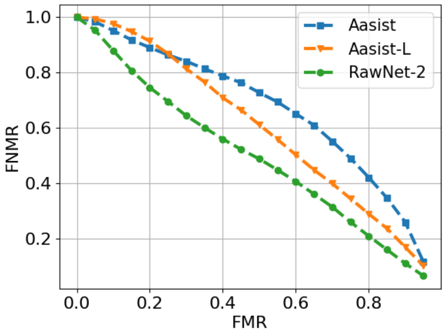

Figure 54: The Detection Error Trade-off (DET) Curve for the Urdu test set created using XTTS-v2. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

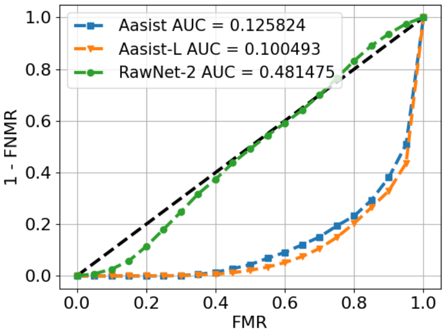

Figure 55: The Receiver Operating Characteristic (ROC) Curve for the Urdu test set created using FreeVC24. Low Area Under the Curve (AUC) values of audio deepfake detection models indicate poor discriminative power of these models on our test set.

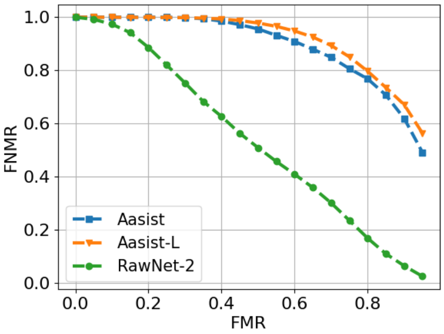

Figure 56: The Detection Error Trade-off (DET) Curve for the Urdu test set created using FreeVC24. The trend of DET curves towards the upper right indicates the poor capability of the audio deepfake detection models in distinguishing between bonafide and synthetic audios.

## C Authenticity of Mimicry Subset

Figures 57–73 illustrate the proximity of the mimicry subset audios with the bonafide IndicSuperb audios for the target languages.

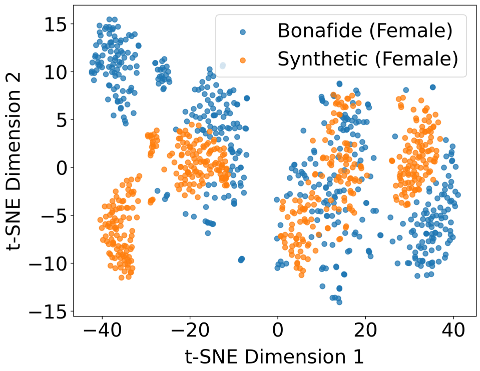

Figure 57: The t-SNE plot of bonafide (IndicSuperb) Gujarati and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

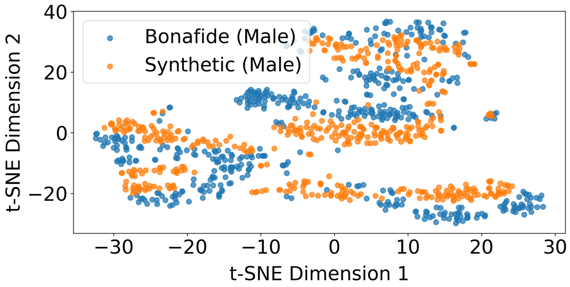

Figure 58: The t-SNE plot of bonafide (IndicSuperb) Gujarati and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

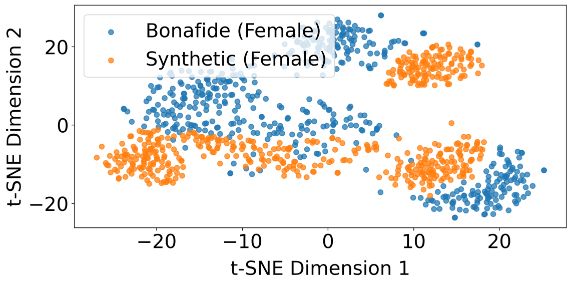

Figure 59: The t-SNE plot of bonafide (IndicSuperb) Kannada and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

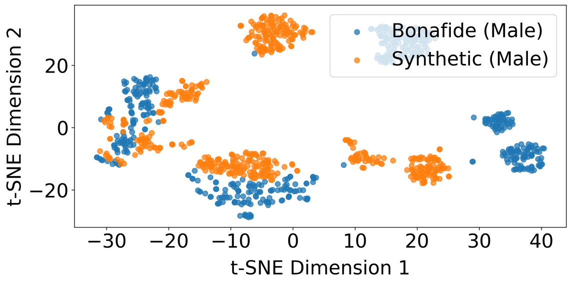

Figure 60: The t-SNE plot of bonafide (IndicSuperb) Kannada and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

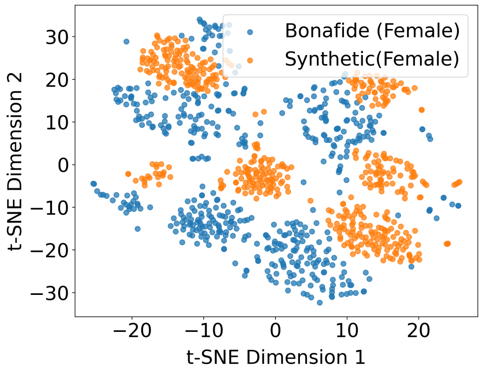

Figure 61: The t-SNE plot of bonafide (IndicSuperb) Malayalam and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

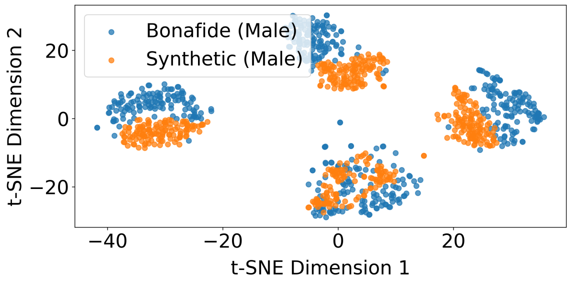

Figure 62: The t-SNE plot of bonafide (IndicSuperb) Malayalam and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

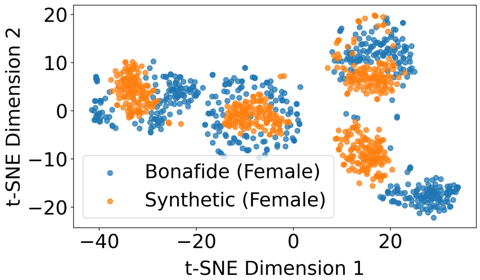

Figure 63: The t-SNE plot of bonafide (IndicSuperb) Marathi and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

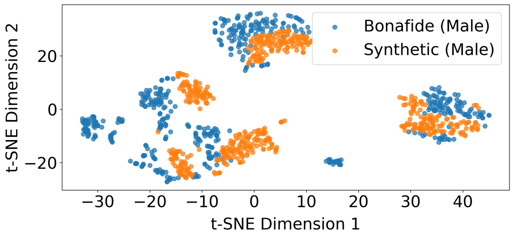

Figure 64: The t-SNE plot of bonafide (IndicSuperb) Marathi and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

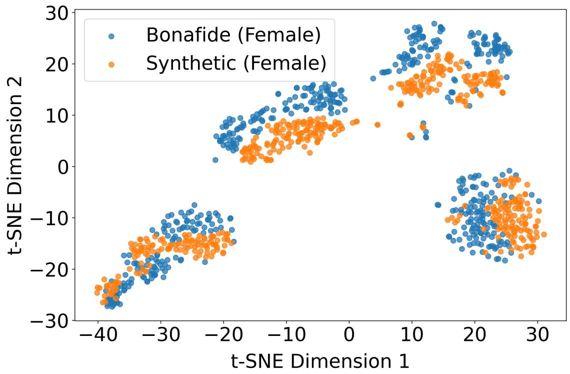

Figure 65: The t-SNE plot of bonafide (IndicSuperb) Punjabi and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

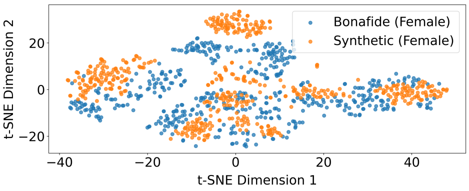

Figure 66: The t-SNE plot of bonafide (IndicSuperb) Punjabi and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

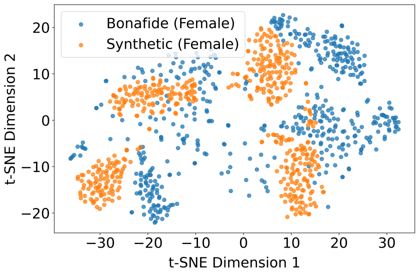

Figure 67: The t-SNE plot of bonafide (IndicSuperb) Tamil voices and IndicSynth's mimicry subset's synthetic clips of the same target female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

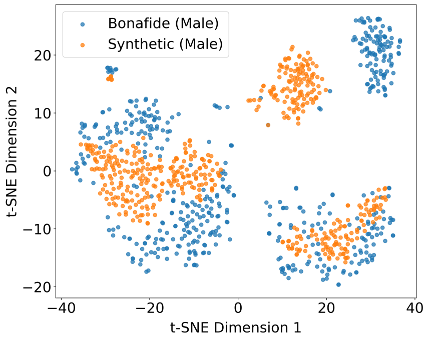

Figure 68: The t-SNE plot of bonafide (IndicSuperb) Tamil voices and IndicSynth's mimicry subset's synthetic clips of the same target male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

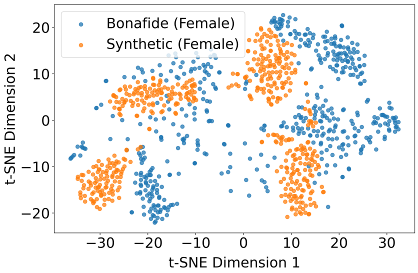

Figure 69: The t-SNE plot of bonafide (IndicSuperb) Tamil voices and IndicSynth's mimicry subset's synthetic clips of the same target female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

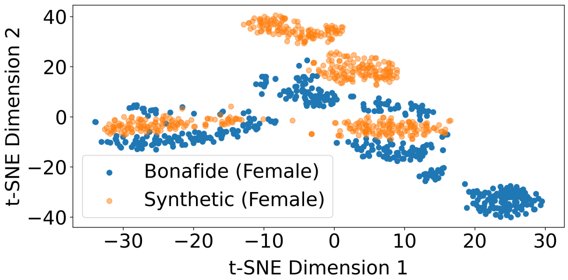

Figure 70: The t-SNE plot of bonafide (IndicSuperb) Telugu and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

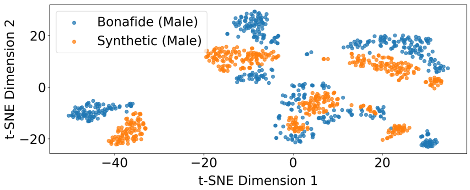

Figure 71: The t-SNE plot of bonafide (IndicSuperb) Telugu and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

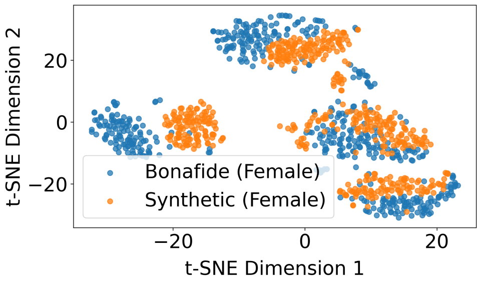

Figure 72: The t-SNE plot of bonafide (IndicSuperb) Urdu and IndicSynth's mimicry subset's female speakers. The plot reveals the proximity of the bonafide and synthetic audios.

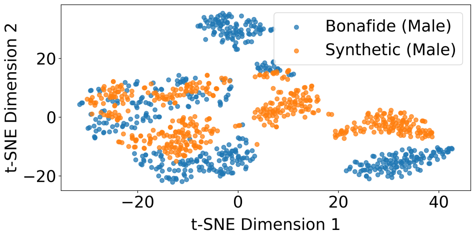

Figure 73: The t-SNE plot of bonafide (IndicSuperb) Urdu and IndicSynth's mimicry subset's male speakers. The plot reveals the proximity of the bonafide and synthetic audios.

## D IndicSynth Linguistic Authenticity: Additional Plots

Figures 74–76 illustrate the t-SNE plots from the embeddings obtained through the language identification model for Sanskrit, Punjabi, and Urdu. The plots demonstrate the linguistic authenticity of IndicSynth's Sanskrit, Punjabi, and Urdu audios.

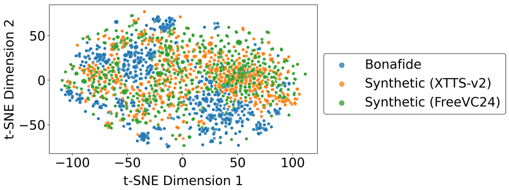

Figure 74: t-SNE visualization of bonafide (IndicSuperb) and synthetic (IndicSynth) Sanskrit dataset. The plot indicates that the IndicSynth-Sanskrit subset has effectively captured Sanskrit's linguistic traits.

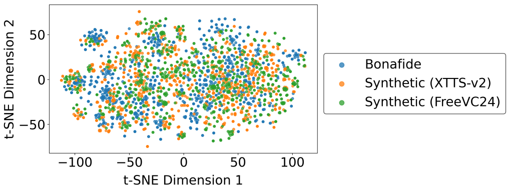

Figure 75: t-SNE visualization of bonafide (IndicSuperb) and synthetic (IndicSynth) Punjabi dataset. The plot indicates that the IndicSynth-Punjabi subset has effectively captured Punjabi's linguistic traits.

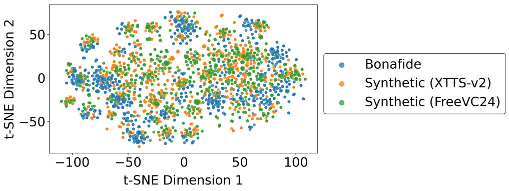

Figure 76: t-SNE visualization of bonafide (IndicSuperb) and synthetic (IndicSynth) Urdu dataset. The plot indicates that the IndicSynth-Urdu subset has effectively captured Urdu's linguistic traits.

## E Costs

This section highlights the costs of various experiments conducted for this study in terms of carbon emissions, electricity consumption, and execution time. The data generation and experimentation were done using the NVIDIA RTX A6000 GPU. Below are the details:

1. **Cost of fine-tuning XTTS-v2:** Fine-tuning the XTTS-v2 for one epoch for a particular target language takes approximately 1.5 hours.

In those 1.5 hours, the process causes approximately 0.403 kgCO2eq carbon emission and consumes 0.564 kWh of electricity. This result is for the training set containing 70,692 audio clips. We fine-tuned each model for 45 epochs. The XTTS-v2 model contains 470,751,571 parameters.

2. **XTTS-v2**: It takes about 45 minutes to generate 1,000 audio clips from the XTTS-v2 model for Hindi. The average duration of a Hindi audio clip in IndicSuperb is 2.65 seconds. For generating 1000 synthetic audios using the XTTS-v2 model, approximately 0.106 kgCO2eq carbon emission and 0.148 kWh of electricity are consumed. The XTTS-v2 model has 470,751,571 parameters and occupies 2,110 MB of memory.
3. **FreeVC24**: It takes about 4 minutes to generate 1000 audio clips from the FreeVC24 model for Hindi. The average duration of a Hindi audio clip in IndicSuperb is 2.65 seconds. For generating 1000 synthetic audios using the FreeVC24 model, approximately 0.010 kgCO2eq carbon emissions and 0.014 kWh of electricity are consumed. The FreeVC24 model contains 356,216,448 parameters and occupies 1690 MB of memory.
4. **VITS**: The VITS model has 83,050,540 parameters and occupies 586 MB of memory.
5. **Language identification**: It took about 2 minutes to run the VoxLingua107 ECAPA-TDNN spoken language identification model on a test set containing 8,000 audio clips. The model caused approximately 0.004 kgCO2eq carbon emission and energy consumption of 0.006 kWh.
6. **ResNet TDNN speaker verification model**: It took about 4 minutes to generate ResNet TDNN embeddings for 10,000 audio clips. The model caused approximately 0.012 kgCO2eq carbon emission and energy consumption of 0.018 kWh for 10,000 embeddings. The ResNet TDNN model has 17,282,816 parameters and occupies 334 MB of memory.
7. **Ecapa-TDNN speaker verification model**: It took about 3 minutes to generate ResNet TDNN embeddings for 10,000 audio clips. The model caused approximately 0.007 kgCO2eq carbon emission and energy consumption of 0.009 kWh for 10,000 embeddings. The Ecapa-TDNN model has 22,150,912 parameters and occupies 364 MB of memory.
8. **X-Vector speaker verification model**: It took about 1.5 minutes to generate ResNet TDNN embeddings for 10,000 audio clips. The model caused approximately 0.003 kgCO2eq carbon emission and energy consumption of 0.004 kWh for 10,000 embeddings. The X-Vector model has 8,172,473 parameters and occupies 300 MB of memory.
9. **AASIST Audio Deepfake Detection Model**: The AASIST model has 297,866 parameters and occupies 264 MB of memory.
10. **AASIST-L Audio Deepfake Detection Model**: The AASIST model has 85,306 parameters and occupies 262 MB of memory.
11. **RawNet-2 Audio DeepFake Detection Model**: The RawNet-2 model contains 17,623,671 parameters and occupies 67.2 MB of disk space.

## F Tools and Software used

We used the following tools and software for this study (other than the ones already cited in the paper):

1. We used Grammarly and ChatGPT for better sentence construction at occasional places and to enhance clarity in our draft.
2. We used draw.io and the matplotlib for diagrams.
3. We used Librosa to generate the MFCC features.
4. We used Pytorch for experimentation: version 2.5.1+cu124.

## G Licenses

In this section, we specify the licenses of the datasets and models that we have used for this study.

1. IndicSuperb: Creative Commons CC0 license ("no rights reserved").
2. Aasist, Aasist-L, and the RawNet-2: The ADD models used are licensed under the MIT License.
3. The speechbrain models (Ecapa-TDNN, ResNet TDNN, X-Vector, VoxLingua107 ECAPA-TDNN spoken language identification model) is licensed under the Apache License 2.0.