GreenVoice: Language-Agnostic, Sustainability-Aware, and Scalable Evaluation of Synthetic Speech Generation Models

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in synthetic speech generation models have enabled highly realistic speech that emulates human voices. These models generate voice clones that closely mimic target speakers. Consequently, they pose privacy and security risks through audio deepfakes and audio spoofs. Audio deepfakes are synthetic speech created to deceive humans, whereas audio spoofs are voice clones created to deceive speaker verification systems. Combating these threats requires robust and generalizable audio deepfake detection (ADD) and anti-spoofing models. However, existing models often exhibit domain-specific biases. Developing robust models requires large-scale multilingual synthetic speech datasets. Most existing datasets are in English or Chinese. Creating new datasets poses three challenges. First, it often involves fine-tuning generation models for target languages, which is expensive. Second, speech quality assessments rely on human evaluations, which lack scalability. Third, large-scale speech generation can incur substantial carbon emissions. To address these gaps, we propose GreenVoice, a language-agnostic, sustainability-aware, and scalable framework for evaluating synthetic speech generation models. We empirically validate GreenVoice through a human evaluation with 92 participants. Using GreenVoice, we conduct evaluations across 10 generation models, two English accents, and 12 Indian languages. Leveraging evaluation insights, we introduce Indic-EcoSynth, a 2100-hour multilingual synthetic speech dataset for ADD and anti-spoofing research.

1 Introduction

Recent advances in synthetic speech generation technologies have enabled the generation of highly realistic speech that emulates human voices (Kawa et al., 2022). These technologies are also capable of producing *voice clones* that closely mimic the bonafide (real) voices of target speakers (Zhu et al., 2024; Ji et al., 2025). These capabilities enable applications across diverse domains, including assistive technologies, education, media, and entertainment (Ahmed et al., 2024; Wang et al., 2025; Popa et al., 2026). Despite socially beneficial applications, these technologies also pose significant risks to privacy, security, and trust in digital communications through *audio deepfakes* and *audio spoofs* (Kumar et al., 2025).

Audio deepfakes are synthetic speech recordings created to deceive humans. Such recordings are often misused to spread misinformation, defame individuals, and trigger public unrest (Müller et al., 2024b; Hutiri et al., 2024; Taleby Ahvanooey et al., 2025). In contrast, audio spoofs are voice clones created to deceive security systems, such as speaker verification (Kawa et al., 2022). Audio spoofing may lead to impersonation and identity theft (Liu et al., 2023). For instance, in 2019, impostors exploited an audio spoof of a corporate executive to orchestrate a financial fraud of 243,000 US dollars (USD) (Munir et al., 2024). Similarly, an audio deepfake-based cybercrime led to a loss of 35 million USD for a UAE-based company (Rabhi et al., 2024). Synthetic speech attacks have increased by over 2000% within a three-year period (Nguyen et al., 2025). Therefore, developing robust *audio deepfake detection (ADD)* and *anti-spoofing* systems is crucial.

Developing such models requires large-scale synthetic speech datasets (Li et al., 2024). However, most available datasets are in English or Chinese (Munir et al., 2024). Consequently, most prior works in ADD and anti-spoofing focus on these languages, leaving several low-resource languages underserved (Ba et al.,

2023; Yi et al., 2024). Moreover, existing ADD and anti-spoofing models exhibit significant domain-specific biases across languages, accents, and generation models (Müller et al., 2024a; Rabhi et al., 2024; Gasenzer & Wolter, 2024). Minor linguistic variations can increase the attack success rate by more than 60% (Nguyen et al., 2025). This limitation is particularly concerning for linguistically diverse countries such as India. India has 22 scheduled languages, 121 major languages, and 1,369 mother tongues spoken by more than one billion speakers (Udupa et al., 2025; Gupta et al., 2025; Tomar et al., 2025). However, Indian languages are underrepresented in mainstream benchmarks (Sharma et al., 2026).

As illustrated in Figure 1, creating large-scale multilingual ADD and anti-spoofing datasets poses three key challenges. First, speech quality assessment primarily relies on human evaluation (Reddy et al., 2022; Varadhan et al., 2025). Such evaluations require participants who are proficient in the target languages. Recruiting a sufficiently large and diverse group of evaluators is particularly challenging for low-resource, geographically dispersed languages. Consequently, human evaluation is difficult to scale across a large number of languages, leaving many languages underrepresented. Second, dataset creation often involves training or fine-tuning synthetic speech generation models for target languages. These processes can be computationally expensive, undermining inclusivity by deterring researchers with limited computational resources from creating datasets (Purohit et al., 2024). Third, large-scale speech generation can cause significant carbon emissions. While the environmental impact of speech generation is underexplored, prior work emphasizes that substantial reductions in carbon emissions are essential to mitigating the increasing frequency of natural disasters (Strubell et al., 2019). Therefore, it is crucial to select generation models responsibly when creating large-scale datasets.

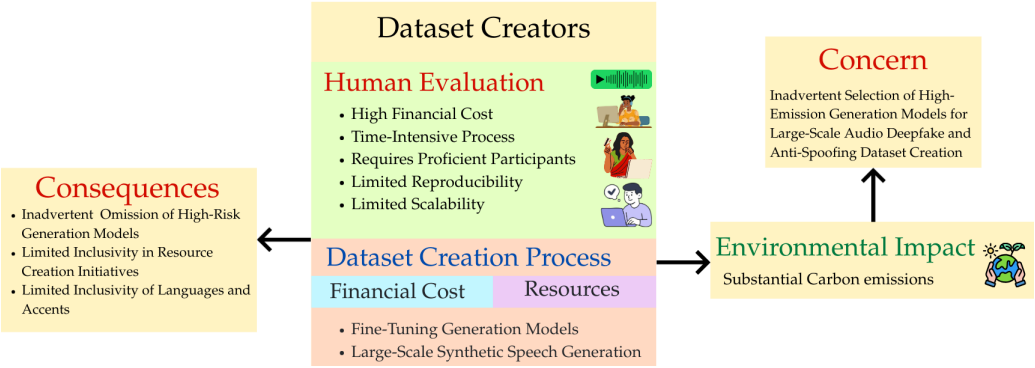

Figure 1: Challenges in creating multilingual synthetic speech datasets for ADD and anti-spoofing research.

Prior ADD and anti-spoofing datasets typically use multiple generation models that support the target languages (Munir et al., 2024; Sharma et al., 2025; Huang et al., 2025; Ciobanu et al., 2026). However, without systematic evaluations, models that produce highly realistic speech and convincing voice clones may be inadvertently omitted from dataset creation. Therefore, there is an urgent need for a scalable, language-agnostic framework for jointly evaluating speech quality and the environmental impact of generation models.

In this work, we introduce *GreenVoice*, a language-agnostic and sustainability-aware framework for large-scale comparative evaluation of synthetic speech generation models. The proposed framework is motivated by the hypothesis that audio deepfake detection (ADD) and speaker verification (SV) models can be used to evaluate the quality of ADD and anti-spoofing datasets. We validate this hypothesis through a large-scale human evaluation with 92 participants and 10 synthetic speech generation models. The framework comprises five automated metrics: Realism Score, Similarity Score, Cloning Risk Assessment, Environmental Impact Assessment, and G-Score. Using GreenVoice, we conduct a large-scale comparative evaluation of 10 synthetic speech generation models across two English accents. We also apply GreenVoice to evaluate three synthetic speech generation models across 12 Indian languages. Based on evaluation insights, we introduce *Indic-EcoSynth*, a large-scale multilingual synthetic speech dataset for ADD and anti-spoofing research. Indic-EcoSynth contains about 2,100 hours of synthetic speech from 989 speakers (456 females and 533 males) across 12 Indian languages.

Key contributions of this study are:

1. We propose *GreenVoice*, a language-agnostic, sustainability-aware, and scalable framework for evaluating synthetic speech generation models. GreenVoice is based on the hypothesis that audio deepfake detection (ADD) and speaker verification (SV) models can be used to evaluate the quality of multilingual ADD and anti-spoofing datasets.
2. To validate this hypothesis, we conduct a large-scale human evaluation with 92 participants and 10 synthetic speech generation models.
3. Using GreenVoice, we conduct a case study evaluating 10 synthetic speech generation models across two English accents.
4. We further apply GreenVoice to evaluate three synthetic speech generation models across 12 Indian languages.
5. Guided by GreenVoice evaluation insights, we introduce *Indic-EcoSynth*, a large-scale multilingual synthetic speech dataset for ADD and anti-spoofing research. Indic-EcoSynth contains about 2,100 hours of synthetic speech from 989 speakers (456 females and 533 males) across 12 Indian languages.

2 Related Works

Data Scarcity: The proliferation of audio deepfake and spoofing-related fraud motivated research communities to organize ADD, ASVspoof, and spoofing-aware speaker verification challenges (Liu et al., 2023; weon Jung et al., 2022; Yi et al., 2023). Despite these initiatives, existing ADD and anti-spoofing models exhibit substantial domain-specific biases across languages, accents, and synthetic speech generation models (Müller et al., 2024a; Rabhi et al., 2024). Bias mitigation often involves domain adaptation or data augmentation during training (Korshunov & Marcel, 2022; Ba et al., 2023). Such approaches require high-quality synthetic speech datasets spanning diverse languages and generation models (Li et al., 2024). However, most publicly available datasets are in English or Chinese (Müller et al., 2024b; Munir et al., 2024). Consequently, several languages remain underserved. Data scarcity continues to hinder ADD and anti-spoofing research for low-resource languages (Sharma et al., 2025).

Existing Metrics and Their Limitations: Creating ADD and anti-spoofing datasets requires rigorous quality evaluations (Munir et al., 2024). Human evaluation is costly, difficult to scale, and lacks reproducibility (Varadhan et al., 2025). Automated mean opinion score predictors often exhibit limited generalizability to out-of-domain datasets (Cooper et al., 2022). Similarly, Bilingual Evaluation Understudy, Mel-Cepstral Distortion, F0-root mean square error, and Perceptual Evaluation of Speech Quality often exhibit weak correlation with human judgments (Salesky et al., 2021; Maiti et al., 2022; Saeki et al., 2024). Notably, Munir et al. (2024) and Sharma et al. (2025) introduced ADD datasets in low-resource languages and demonstrated the vulnerability of ADD and SV models to synthetic speech in unseen languages. Furthermore, environmental impact assessments are uncommon when generating datasets at scale. Pareto-optimal solutions evaluate the trade-off between speech quality and environmental impact (Douwes et al., 2021; Passoni et al., 2025). However, they often lack scalability. Therefore, there is a pressing need for a scalable, language-agnostic framework for jointly evaluating synthetic speech quality and the environmental impact of generation models. To address this gap, we propose GreenVoice.

3 GreenVoice

This study introduces GreenVoice, a language-agnostic and sustainability-aware framework for evaluating synthetic speech generation models. The framework includes five automated metrics: Realism Score, Similarity Score, Environmental Impact Assessment (EIA) Score, Cloning Risk Assessment (CRA), and G-Score. This section describes GreenVoice and its hypothesis.

Hypothesis: Synthetic speech generation models that produce highly realistic utterances capable of deceiving audio deepfake detection (ADD) systems pose a heightened risk of native-language-based audio deepfake attacks. Similarly, models that generate voice clones capable of deceiving speaker verification (SV) systems

pose a heightened risk of multilingual audio spoofing attacks. Identifying these high-risk generation models is crucial for developing robust audio deepfake detection and anti-spoofing systems.

Equal error rate (EER) is a widely used metric for evaluating ADD and SV models (Yi et al., 2022; weon Jung et al., 2022). While EER is traditionally used to assess the performance of ADD and SV models, we hypothesize that the metric is also valuable for evaluating the risks associated with synthetic speech generation models. For instance, a higher EER in ADD indicates that the ADD model struggles to distinguish between bonafide and synthetic speech. It further suggests that the generation model produces highly realistic synthetic speech that is challenging for ADD models to detect. Similarly, when an SV model exhibits higher EER on an audio spoofing test set, it highlights the vulnerability of the SV model to such attacks. It also suggests that the voice clones produced by the generation model exhibit substantial proximity to the bonafide target voices.

In summary,

1. An elevated EER in ADD indicates that the synthetic speech produced by the generation model exhibits high realism.
2. An elevated EER in SV indicates that the voice clones produced by the generation model exhibit substantial proximity to bonafide target voices.

Based on this hypothesis, GreenVoice evaluates synthetic speech generation models using multiple risk- and sustainability-oriented metrics. As illustrated in Figure 2, suppose we have T synthetic speech generation models for evaluation using GreenVoice. We first compute the Realism and Similarity Scores of these models using ADD and SV models as evaluators.

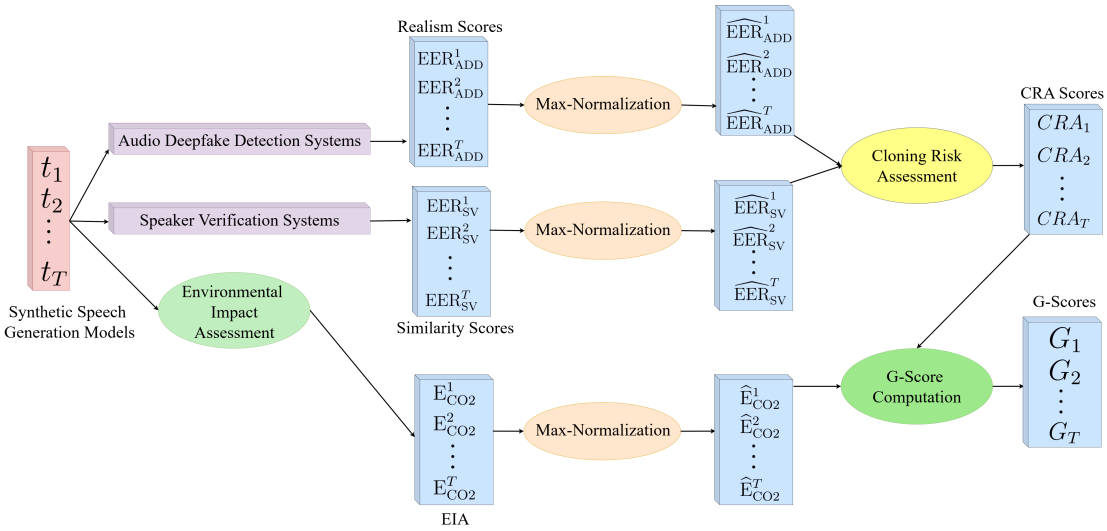

Figure 2: GreenVoice framework for automated, large-scale, multi-objective comparative evaluation of T synthetic speech generation models. Realism and Similarity Scores are combined to compute Cloning Risk Assessment (CRA), which is further integrated with Environmental Impact Assessment (EIA) scores to compute G-Scores.

Realism Score: The Realism Score measures the authenticity of speech produced by a synthetic speech generation model. To compute the Realism Score of a generation model, we first prepare a test set containing p voice samples. This set includes an equal number of bonafide and synthetic clips produced by that generation model. Next, we feed these test samples to m ADD models and compute EER values with each of these models. The averaged EER across all the evaluator ADD models represents the *Realism Score* of the target synthetic speech generation model.

Formally, the Realism Score of a synthetic speech generation model is defined as:

$$\text{EER}_{\text{ADD}} = \frac{1}{m}\sum_{i=1}^{m}\text{EER}_{\text{ADD}}^{(i)}, \tag{1}$$

where $\text{EER}_{\text{ADD}}^{(i)}$ is the equal error rate obtained from the i-th ADD model, and m is the total number of evaluator ADD models.

Realism Score reflects the vulnerability of ADD models to synthetic speech produced by the target generation model. Consequently, the metric identifies generation models that pose an elevated risk of native-language audio deepfakes. Such models can be used to create challenging ADD datasets.

Similarity Score: The Similarity Score measures the proximity of the bonafide target voices to the voice clones produced by the generation model. The score computation process employs SV models as similarity evaluators. SV models accept two speech recordings as input and determine if they belong to the same speaker. The input speech recordings form a trial pair.

To compute the Similarity Score of a synthetic speech generation model, we first create a test set of q trial pairs, with equal numbers of positive and negative pairs. A positive trial pair contains two bonafide (real) voice samples of speaker X. In contrast, a negative trial pair consists of a bonafide voice sample of speaker X and a cloned voice sample of X. Next, we feed these trial pairs to k speaker verification models and compute EER values with each of these models. The averaged EER across all evaluator SV models represents the *Similarity Score* (EER_{SV}) of the generation model.

Formally, the Similarity Score of a synthetic speech generation model is defined as:

$$\text{EER}_{\text{SV}} = \frac{1}{k}\sum_{j=1}^{k}\text{EER}_{\text{SV}}^{(j)}, \tag{2}$$

where $\text{EER}_{\text{SV}}^{(j)}$ denotes the equal error rate obtained using the j-th SV model, and k denotes the number of evaluator SV models.

Similarity Score identifies generation models that pose an elevated risk of multilingual audio spoofs. Such models can be explored for creating challenging anti-spoofing datasets.

Max-Normalization: The Realism (EER_{ADD}) and Similarity (EER_{SV}) Scores are distinct metrics that differ in scale and capture different aspects of the model. Therefore, to map these metrics to a fixed range $[0, 1]$ and enable unified comparative evaluation, we max-normalize these scores. Max-normalization is performed by dividing each score by the maximum value of that score across all target generation models, as follows:

$$\widehat{\text{EER}}_{\text{ADD}}^{i} = \frac{\text{EER}_{\text{ADD}}^{i}}{\max\limits_{j=1,\ldots,T}\text{EER}_{\text{ADD}}^{j}} \tag{3}$$

$$\widehat{\text{EER}}_{\text{SV}}^{i} = \frac{\text{EER}_{\text{SV}}^{i}}{\max\limits_{j=1,\ldots,T}\text{EER}_{\text{SV}}^{j}} \tag{4}$$

The resulting max-normalized scores reflect the relative performance of the target generation models with respect to realism and similarity. Consequently, the max-normalized Realism and Similarity scores become directly comparable and suitable for integration into a single composite evaluation metric, namely CRA.

Cloning Risk Assessment: The Cloning Risk Assessment (CRA) Score of a synthetic speech generation model t_i is the weighted average of max-normalized Realism and Similarity Scores of that model (t_i), as illustrated below:

$$\text{CRA}_i = \frac{w_{\text{ADD}} \times \widehat{\text{EER}}_{\text{ADD}}^{i} + w_{\text{SV}} \times \widehat{\text{EER}}_{\text{SV}}^{i}}{w_{\text{ADD}} + w_{\text{SV}}} \tag{5}$$

Here, w_{ADD} and w_{SV} are weights given to realism and similarity dimensions.

Environmental Impact Assessment: The next step involves computing the Environmental Impact Assessment (EIA) scores of these models. EIA of a generation model t_i is the measure of carbon emissions from generating c synthetic speech recordings using t_i. This metric is mathematically denoted as $\mathrm{E}_{\mathrm{CO2}}$. Similar to the Realism and Similarity Scores, we max-normalize EIA as follows:

$$\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}} = \frac{\mathrm{E}^{i}_{\mathrm{CO2}}}{\max\limits_{j=1,\ldots,T} \mathrm{E}^{j}_{\mathrm{CO2}}} \tag{6}$$

G-Score: The G-Score of a generation model is a unified composite metric that integrates the CRA and max-normalized EIA scores ($\widehat{\mathrm{E}}_{\mathrm{CO2}}$) of that model. CRA serves as a performance metric, whereas $\widehat{\mathrm{E}}_{\mathrm{CO2}}$ serves as a sustainability metric.

Formally, we compute the G-Score of a generation model t_i as illustrated below:

$$\textit{G-Score}_i = \frac{\mathrm{CRA}_i}{\left(\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}}\right)^{\frac{1}{n}}} \tag{7}$$

Here, n is the scaling factor that controls the influence of EIA ($\widehat{\mathrm{E}}_{\mathrm{CO2}}$) on G-Scores.

A higher G-Score indicates that the synthetic speech generation model exhibits higher cloning risk while incurring lower inference-time carbon emissions. Such models can be leveraged to create large-scale multilingual ADD and anti-spoofing datasets with reduced environmental impact.

4 Hypothesis Validation

We validate the GreenVoice hypothesis through a human evaluation with 92 participants[1] across 10 synthetic speech generation models. This section presents the correlation between objective Realism and Similarity Scores and human judgments.

4.1 Setup

Accent, Dataset, and Generation Models: Since participants were proficient speakers of Indian English, the subjective evaluation was conducted on Indian English speech. For bonafide Indian English speech, we use the NISP dataset (Kalluri et al., 2021). The evaluation session lasted approximately 30 minutes. Participants received an honorarium of INR 150 for their participation. The study was conducted across 10 publicly available generation models: CosyVoice 2, FishTTS, MetaVoice-1B, OpenVoice V2, SV2TTS, XTTS-v2, YourTTS, FreeVC, DiffVC, and SeedVC (Jia et al., 2018; Popov et al., 2022; Casanova et al., 2022; Li et al., 2023; Du et al., 2024; Metavoiceio, 2024; Casanova et al., 2024; Qin et al., 2024; Liao et al., 2024; Liu, 2024).

Realism and Similarity Score Computation: We used the AASIST and RawNet-2 as evaluators for Realism Score (Jung et al., 2022; Tak et al., 2021). Similarly, we used ECAPA-TDNN and ResNet-TDNN as evaluators for Similarity Score (Ravanelli et al., 2021; Desplanques et al., 2020; Villalba et al., 2020). For each synthetic speech generation model, the Realism Score test set includes 4,000 voice samples ($p = 4000$), and the Similarity Score test set contains 20,000 trial pairs ($q = 20,000$). We followed the score computation process described in Section 3.

Human Evaluation for Realism: We created a test set comprising 22 randomly selected samples. Of these, two were bonafide speech recordings, corresponding to one female and one male speaker. The remaining 20 samples were synthetic speech recordings, with two samples from each of the 10 target generation models. For each model, the synthetic samples consisted of one female and one male voice. Accordingly, we define 11 *categories*, corresponding to bonafide speech and synthetic speech from the 10 generation models.

[1]Our human evaluation was reviewed and approved by the Institutional Review Board (IRB). We also obtained informed consent from participants. Participation was voluntary, participants were adults, and responses were anonymized.

Participants were allowed to listen to a test sample as many times as needed. After listening to a test sample, participants were required to answer the following question via a Google Form: *How realistic (human-like) does the test audio sample sound?* Participants rated each sample on a 5-point Likert scale (1 = entirely synthetic, 5 = completely human-like). Subsequently, for each category, we compute the mean opinion score (MOS) by averaging the ratings of all test samples in that category across all participants. Consequently, we get a single MOS score per category for realism.

Human Evaluation for Similarity: We created a test set comprising 20 randomly selected trial pairs. A trial pair consisted of one bonafide speech recording of speaker X and one synthetic speech recording (voice clone) of the same speaker X. The test set was constructed by selecting two trial pairs for each of the 10 generation models. These two trial pairs consisted of one female-speaker pair and one male-speaker pair. Participants were allowed to listen to each trial pair multiple times.

After listening to a trial pair, participants were required to answer the following question via a Google Form: *How similar do the two voices in the trial pair sound?* Participants rated each trial pair on a 5-point Likert scale (1 = completely different voices, 5 = exactly the same voice). Subsequently, for each target generation model, we compute the mean opinion score (MOS) by averaging the ratings of all corresponding trial pairs across all participants. Consequently, we get a single MOS score per model for similarity.

4.2 Correlation Analysis

Realism Score with Subjective MOS: Table 1 illustrates the objective and subjective Realism Scores of each target generation model. The bonafide speech recordings achieved an MOS of only 3.86, which suggests an upper bound on perceived realism. Taking this into consideration, we focus our analysis on generation models that attained a subjective MOS of 3 or higher. *FreeVC24*, *OpenVoice V2*, and *SeedVC* consistently exhibited elevated Realism Scores and MOS. However, FishTTS and MetaVoice-1B attained high MOS but low objective Realism Scores. Similarly, DiffVC attained a low MOS but a high objective Realism Score. Furthermore, Spearman's rank correlation analysis reveals a moderate positive correlation between the objective Realism Scores and human mean opinion scores ($\rho = 0.59$, p-value $= 0.07$), suggesting that higher Realism Scores tend to correspond to higher perceived realism. The correlation is marginally significant.[2]

Table 1: Objective and subjective Realism Scores for target synthetic speech generation models. Subjective MOS scores ≥ 3 and objective realism scores ≥ 45 are shown in bold and underlined. *FreeVC24*, *OpenVoice V2*, and *SeedVC* show consistently high realism across both objective and subjective assessments. Spearman's Rank Correlation: $\rho = 0.59$, p-value $= 0.07$.

Gen. Model	Obj. Realism Score	Subj. MOS
Bonafide	–	**3.86**
CosyVoice 2	34.91	2.88
DiffVC	**66.63**	2.84
FishTTS	18.46	**3.68**
FreeVC24	**72.94**	**4.33**
MetaVoice-1B	23.96	**3.21**
OpenVoice V2	**46.17**	**3.20**
SeedVC	**46.95**	**3.55**
SV2TTS	8.87	2.12
XTTS-v2	9.74	2.85
YourTTS	2.97	2.08

Similarity Score with Subjective MOS: Table 2 illustrates the objective and subjective Similarity Scores of each target generation model. *SeedVC* consistently exhibited elevated Similarity and MOS scores. However, FishTTS and FreeVC24 attained high MOS scores but low Similarity Scores. Similarly, CosyVoice 2 attained a low MOS score but a high Similarity Score. Nevertheless, Spearman's rank correlation analysis

[2]Audio samples used in the subjective Realism MOS assessment are included in the supplementary material.

reveals a strong and statistically significant correlation between the objective Similarity Score and human judgments ($\rho = 0.75$, p-value $= 0.01$), indicating that the objective Similarity Score reliably reflects the perceived similarity between generated and target voices.[3]

Table 2: Objective and subjective Similarity Scores of target synthetic speech generation models. Subjective MOS scores ≥ 3 and objective Similarity Scores ≥ 25 are shown in bold and underlined. *SeedVC* consistently exhibits a relatively high similarity score in both objective and subjective evaluations. Spearman's Rank Correlation: $\rho = 0.75$, p-value $= 0.01$.

Gen. Model	Obj. Similarity Score	Subj. MOS
CosyVoice 2	**27.18**	2.59
DiffVC	0.99	1.48
FishTTS	8.76	**3.84**
FreeVC24	1.96	**3.27**
MetaVoice-1B	4.64	2.01
OpenVoice-V2	2.07	2.57
SeedVC	**25.77**	**3.43**
SV2TTS	0.62	1.43
XTTS-v2	3.92	2.11
YourTTS	0.76	1.70

5 Utility of GreenVoice

5.1 GreenVoice Evaluation Across Accents

With the continuous emergence of new synthetic speech generation models, there is also an urgent need for systematic evaluation and identification of high-risk generation models. Such identification is a critical first step toward mitigating the threats posed by these models. Therefore, we explore how GreenVoice can assist responsible AI researchers in identifying high-risk models. We present a case study using GreenVoice to evaluate 10 generation models across two English accents: UK–US English and Indian English.

Setup: We use the same dataset, generation models, ADD and SV evaluators, and score computation process as described in Section 4.1. We obtain bonafide UK–US English speech from LibriTTS (Zen et al., 2019). Since the target generation models are primarily trained on UK–US English, UK–US English is the *in-domain* accent and Indian English is the *out-of-domain* accent. Table 3 illustrates the scores of the target generation models across two accents for each GreenVoice metric.[4] Below is a metric-wise assessment.

Realism Score: On the in-domain accent (UK–US English), CostVoice 2 (CV2) and FishTTS (FT) exhibit high Realism Scores. However, their scores decline significantly on the out-of-domain accent (Indian English). In contrast, FreeVC24 (FVC), DiffVC (DVC), and SeedVC (SVC) attain the elevated Realism Scores on out-of-domain accent. DVC consistently attains elevated Realism Scores on both accents. Notably, Realism Scores of most text-to-speech (TTS) models decline substantially under out-of-domain accent. In contrast, voice conversion models (DVC, FVC, and SVC) exhibit greater stability to accent variations.

Similarity Score: CosyVoice 2 (CV2) and SeedVC (SVC) consistently exhibit elevated Similarity Scores across both in-domain and out-of-domain accents. Notably, all the target generation models consistently attain elevated scores on in-domain accent as compared to out-of-domain.

Realism-Similarity Trade-off: While DiffVC (DVC) and FreeVC24 (FVC) exhibit high Realism Scores, these models attain low Similarity Scores. In contrast, CosyVoice 2 (CV2) exhibits high Similarity Scores across both in- and out-of-domain accents. However, the model attains a low Realism Score on out-of-domain accent. These observations indicate a trade-off between Realism and Similarity Scores of synthetic speech generation models. Generation models that attain high Realism Scores often exhibit low Similarity Scores and vice versa.

[3]Trial pairs for the subjective Similarity MOS assessment are included in the supplementary material.
[4]For brevity, normalized scores are omitted from Table 3.

Table 3: GreenVoice scores of the target synthetic speech generation models across the UK–UK English (in-domain) and Indian English (out-of-domain) accents. For each metric and accent, the highest score is shown in bold and underlined, while the second- and third-highest scores are underlined. In contrast, models with minimal EIA are highlighted. CosyVoice 2 (CV2), FishTTS (FT), MetaVoice-1B (MV), OpenVoice V2 (OV2), SV2TTS (SV2), XTTS-v2 (XV2), YourTTS (YT), DiffVC (DVC), FreeVC24 (FVC), SeedVC (SVC).

Score	Accent	CV2	FT	MV	OV2	SV2	XV2	YT	DVC	FVC	SVC
Realism Score	UK–US	**84.46**	59.18	42.37	45.93	47.92	32.36	32.16	49.71	39.38	44.17
	Indian	34.91	18.46	23.96	46.17	8.87	9.74	2.97	66.63	**72.94**	46.95
Similarity Score	UK–US	**44.60**	29.58	34.15	10.59	17.33	32.49	25.45	8.28	12.30	37.53
	Indian	**27.18**	8.76	4.64	2.07	0.62	3.92	0.76	0.99	1.96	25.77
EIA	–	0.340	1.151	1.794	0.066	0.493	0.166	**0.004**	0.154	0.010	0.302
CRA (Scenario–A)	UK–US	**1.00**	0.68	0.63	0.39	0.48	0.56	0.48	0.39	0.37	0.68
	Indian	0.74	0.29	0.25	0.35	0.07	0.14	0.03	0.47	0.54	**0.80**
CRA (Scenario–B)	UK–US	**1.00**	0.69	0.55	0.48	0.53	0.45	0.42	0.51	0.43	0.59
	Indian	0.58	0.27	0.30	0.52	0.10	0.14	0.04	0.74	**0.81**	0.70
CRA (Scenario–C)	UK–US	**1.00**	0.67	0.71	0.30	0.42	0.66	0.53	0.27	0.31	0.78
	Indian	**0.90**	0.31	0.20	0.19	0.04	0.14	0.03	0.21	0.26	0.89
G-Score (Scenario–A)	UK–US	1.52	0.76	0.63	0.89	0.66	1.01	**2.19**	0.72	1.36	1.06
	Indian	1.12	0.32	0.25	0.81	0.10	0.25	0.16	0.88	**1.96**	1.24
G-Score (Scenario–B)	UK–US	1.52	0.77	0.55	1.10	0.73	0.82	**1.93**	0.94	1.57	0.92
	Indian	0.88	0.30	0.30	1.19	0.14	0.25	0.18	1.36	**2.98**	1.10
G-Score (Scenario–C)	UK–US	1.52	0.75	0.71	0.68	0.59	1.20	**2.45**	0.49	1.15	1.21
	Indian	1.36	0.34	0.20	0.43	0.06	0.26	0.14	0.39	0.94	**1.38**

Cloning Risk Assessment (CRA): We compute the CRA across three scenarios: A, B, and C. *Scenario–A* equally prioritizes the realism and similarity aspects by setting both w_{ADD} and w_{SV} to 1 in CRA computation. This scenario represents the case where the aim is to create a dual-use dataset for ADD and anti-spoofing tasks. CosyVoice 2 (CV2) and SeedVC (SVC) exhibit elevated CRA, offering a balance between realism and similarity. Further, *Scenario–B* prioritizes realism over similarity by setting $w_{\text{ADD}} = 4$ and $w_{\text{SV}} = 1$. This scenario represents the case in which the aim is to create an ADD-focused dataset. ADD-focused datasets typically require a diverse set of highly realistic synthetic voices. SVC exhibits elevated CRA on both accents. Among other models, CV2 and FishTTS (FT) attain elevated CRA on in-domain accent, whereas FreeVC24 (FVC) and DiffVC (DVC) exhibit elevated CRA on out-of-domain accent. In contrast, *Scenario–C* prioritizes similarity over realism by setting $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 4$. This scenario represents the case in which the aim is to create an anti-spoofing dataset. Such datasets require voice clones that closely mimic the bonafide target voices. CV2 and SVC exhibit elevated CRA under this scenario. Importantly, across the three scenarios, TTS models mostly experience a decline in CRA on out-of-domain accent. However, voice conversion models exhibit a high stability to accent variations.

Environmental Impact Assessment (EIA): To compute the EIA of a target model, we tracked the carbon emissions from generating 1,000 ($c = 1,000$) samples from that model (Courty et al., 2024). FreeVC24, OpenVoice V2, and YourTTS exhibit *minimal* carbon emissions. Similarly, DiffVC and XTTS-v2 cause *low* carbon emissions. Furthermore, CosyVoice 2, SeedVC, and SV2TTS cause *moderate* carbon emissions. However, FishTTS and MetaVoice-1B cause *substantial* carbon emissions among these models.

G-Score: We extend the CRA computations to G-Score computations across the three Scenarios: A, B, and C. Here, we set the scaling factor n to 4.[5] In Scenario–A and Scenario–B, YourTTS (YT) attains the highest G-Score for in-domain accent, whereas FreeVC24 (FVC) attains the highest G-Score on out-of-domain accent. In contrast, when similarity is prioritized over realism (Scenario–C), SeedVC (SVC) exhibits the highest G-Score on out-of-domain accent. YT still attains the highest in-domain G-Score across the three scenarios.

Key Insights: *YourTTS (YT)* exhibits the highest in-domain G-Score. One possible reason for its low EIA could be that YT is a zero-shot TTS model built on the VITS and uses HiFi-GAN v1 as the vocoder (Kim et al., 2021). It directly takes raw text as input, thus eliminating the need for grapheme-to-phoneme converters. In contrast, *CosyVoice 2 (CV2)* achieves the highest scores on UK–US English when evaluated

[5]Section 5.2 discusses the impact of scaling factor.

using GreenVoice metrics, excluding G-Scores and EIA. It is reasonable as the generation model is built on a large language model (LLM) backbone. The use of an LLM backbone increases the inference cost of CV2, leading to higher carbon emissions. In Indian English, CV2 attains the highest scores on metrics that prioritize similarity. It could be because CV2 separates the semantic and acoustic information of speech signals and models them independently (Du et al., 2024). Its flow-matching model leverages speaker embeddings to condition audio generation. *SeedVC (SVC)* mostly exhibits elevated scores on evaluation using metrics that value the similarity aspect. It is reasonable as the model applies an external timbre shifter during training to align training conditions with inference (Liu, 2024). Notably, YourTTS and *FreeVC24* both leverage the *VITS* framework, suggesting that synthetic speech from VITS-based models tends to have higher realism. Overall, such large-scale, systematic comparative evaluations of generation models can provide key insights into their behavioral patterns across diverse real-world scenarios. These insights can motivate architectural investigations of high-risk models and the development of more robust countermeasures.

5.2 Scaling Factor Impact

We next investigate the impact of the scaling factor (n) on G-Score outcomes. As illustrated in Table 4, when $n = 1$, models with minimal EIA[6], namely FreeVC24, OpenVoice V2, and YourTTS, attain elevated G-Scores. When $n = 2$, FreeVC24, OpenVoice V2, and SeedVC exhibit elevated G-Scores, despite SeedVC having a moderate EIA. For $n = 4$ and $n = 8$, FreeVC24, SeedVC, and CosyVoice 2 attain elevated G-Scores. SeedVC and CosyVoice 2 have moderate EIA. Furthermore, YourTTS, which has the lowest EIA among all models, attains an elevated G-Score when $n = 1$. However, YourTTS exhibits the lowest G-Score when $n = 8$. Notably, increasing n sharply decreases the standard deviation of G-Scores across generation models. Overall, larger values of n reduce the influence of EIA and place greater emphasis on CRA in G-Score.

Table 4: Impact of the Scaling Factor (n) on G-Score for Indian English ($w_{\text{ADD}} = w_{\text{SV}} = 1$).

Gen. Model	$n = 1$	$n = 2$	$n = 4$	$n = 8$
CosyVoice 2	3.90	1.70	1.12	0.91
DiffVC	5.57	1.63	0.88	0.65
FishTTS	0.45	0.36	0.32	0.30
FreeVC24	**96.17**	**7.18**	**1.96**	**1.03**
MetaVoice-1B	0.25	0.25	0.25	0.25
OpenVoice V2	9.60	1.84	0.81	0.53
SeedVC	4.74	1.94	1.24	1.00
SV2TTS	0.25	0.13	0.10	0.08
XTTS-v2	1.48	0.45	0.25	0.18
YourTTS	15.40	0.73	0.16	0.07
Std. Dev.	**29.34**	**2.08**	**0.61**	**0.38**

5.3 Multilingual Dataset Creation

Next, we apply GreenVoice to evaluate synthetic speech generation models across 12 Indian languages. Leveraging evaluation insights, we create *Indic-EcoSynth*, a large-scale synthetic speech dataset to support multilingual audio deepfake detection (ADD) and anti-spoofing research.

Setup: We compare SeedVC (SVC)-generated speech[7] with the publicly available IndicSynth dataset (Sharma et al., 2025). IndicSynth includes FreeVC24 (FVC) and XTTS-v2 (XV2) generated speech across 12 Indian languages, along with VITS-generated Bengali speech. We followed the IndicSynth generation methodology proposed by Sharma et al. (2025) and applied SeedVC to the bonafide IndicSUPERB dataset to generate synthetic speech (Javed et al., 2023). For each language and generation model, we create separate Realism test sets with 16,000 samples (p = 16,000) and Similarity test sets with 20,000 trial pairs (q = 20,000). The evaluator models remain the same as Section 4.1.

[6]EIA details are in Table 3.

[7]SVC exhibited elevated scores on Indian accent, especially G-Score Scenario-A and Scenario-C (Section 5.1).

Table 5: GreenVoice evaluation of FreeVC24 (FVC)-, VITS-, XTTS-v2 (XV2)-, and SeedVC (SVC)-generated speech across 12 Indian languages. SVC exhibits the highest CRA. Language identification results indicate preservation of target-language linguistic traits. Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Odia (or), Punjabi (pa), Sanskrit (sa), Tamil (ta), Telugu (te), Urdu (ur).

Score	Model	bn	gu	hi	kn	ml	mr	or	pa	sa	ta	te	ur
Realism Score	FVC	**68.58**	**68.70**	**64.39**	**60.85**	**70.56**	**64.74**	**62.78**	**66.26**	**71.34**	**63.21**	**63.79**	**64.26**
	SVC	61.62	62.63	58.36	53.86	58.85	55.92	60.95	59.14	57.86	56.11	55.81	58.09
	VITS	67.99	–	–	–	–	–	–	–	–	-	–	–
	XV2	60.85	56.49	27.45	47.76	57.52	50.28	54.29	50.86	21.15	51.73	50.04	56.53
Similarity Score	FVC	4.55	6.39	7.41	4.18	4.68	6.52	5.10	6.98	2.59	7.87	5.02	6.22
	SVC	**32.42**	**34.93**	**34.34**	**30.65**	**35.90**	**32.82**	**31.68**	**33.94**	**32.03**	**32.07**	**33.29**	**33.61**
	VITS	2.86	–	–	–	–	–	–	–	–	–	–	–
	XV2	25.92	26.45	24.11	21.93	29.21	23.70	23.55	27.86	19.43	24.50	26.46	27.38
CRA	FVC	0.57	0.59	0.61	0.57	0.57	0.60	0.58	0.60	0.54	0.62	0.58	0.59
	SVC	**0.95**	**0.96**	**0.95**	**0.94**	**0.92**	**0.93**	**0.99**	**0.95**	**0.91**	**0.94**	**0.94**	**0.95**
	VITS	0.54	–	–	–	–	–	–	–	–	–	–	–
	XV2	0.84	0.79	0.56	0.75	0.81	0.75	0.80	0.79	0.45	0.79	0.79	0.85
Language ID	Real	89.93	98.61	92.25	88.54	97.43	94.90	–	78.39	41.38	97.50	98.63	39.88
	SVC	84.23	93.44	86.15	82.99	79.91	80.01	–	71.50	37.63	96.14	96.95	44.04
	Δ	5.70	5.17	6.10	5.55	17.52	14.89	–	6.89	3.75	1.36	1.68	**-4.16**

Observations: Table 5 illustrates that FVC consistently achieves the highest Realism Scores across the 12 languages. However, SVC consistently achieves higher scores than XV2. Notably, the IndicSynth creation process involved fine-tuning XV2 (Sharma et al., 2025). In contrast, we did not train or fine-tune SVC for these languages. SVC was evaluated directly without domain adaptation. Nevertheless, SVC consistently achieves the highest Similarity Scores across the 12 languages. Furthermore, SVC consistently exhibits the highest CRA across the 12 languages, when $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 1$.[8]

Linguistic Authenticity Assessment: We evaluate whether SVC-generated speech captures linguistic traits of target languages using the VoxLingua107 ECAPA-TDNN spoken language identification model (Valk & Alumäe, 2021; Sharma et al., 2025). We construct bonafide (IndicSUPERB) and SVC-generated synthetic test sets for each language. Each set contains 8,000 samples. We observe that the accuracy differences (Δ) between the bonafide (real) and SVC sets are mostly below 10%. For Malayalam and Marathi, SVC shows a slightly higher difference but still achieves around 80% accuracy. Notably, the SVC-generated Urdu set achieves higher language identification accuracy than the bonafide Urdu set.[9]

Indic-EcoSynth: We introduce *Indic-EcoSynth*, a SeedVC-generated synthetic speech dataset created to support both multilingual ADD and anti-spoofing research.[10] Indic-EcoSynth includes approximately 2,100 hours of synthetic speech, comprising 1,020,384 audio clips across 12 Indian languages from 989 speakers.[11]

6 Conclusions and Future Work

This paper introduces GreenVoice, a language-agnostic, sustainability-aware, and scalable framework for evaluating synthetic speech generation models. The framework is motivated by the hypothesis that equal error rates from audio deepfake detection (ADD) and speaker verification (SV) systems reflect the risks posed by synthetic speech generation models. Based on this hypothesis, GreenVoice introduces five automated evaluation metrics: Realism Score, Similarity Score, Cloning Risk Assessment, Environmental Impact Assessment, and G-Score. Human evaluation with 92 participants across 10 synthetic speech generation models provides evidence supporting our hypothesis. Using GreenVoice, we conduct evaluations covering 10 synthetic speech generation models, two English accents, and 12 Indian languages. Experimental results indicate a trade-off between Realism and Similarity Scores. Voice conversion models exhibit high stability to

[8]We do not include EIA and G-Score-based comparisons in this analysis, as IndicSynth is an existing dataset. EIA and G-Scores are designed to identify sustainable generation models for constructing new datasets at scale. We use CRA for performance-based comparisons with existing resources.

[9]t-SNE plot for Odia (Or) is in Appendix B.

[10]Corresponding bonafide audios are in IndicSUPERB.

[11]We ensured that the source and target speakers were of the same gender. Dataset statistics are in the Appendix A.

linguistic and accent variations, particularly SeedVC. Based on these findings, we introduce Indic-EcoSynth, a large-scale SeedVC-generated dataset to facilitate multilingual ADD and anti-spoofing research. Indic-EcoSynth contains approximately 2,100 hours of audio from 989 speakers across 12 Indian languages.[12]

This work opens up several avenues for future research. First, GreenVoice enables large-scale behavioral analyses of synthetic speech generation models across diverse languages and accents. Insights from such analyses can motivate architectural investigations of high-risk models and the development of more robust countermeasures. Second, the high-risk models identified through GreenVoice can be leveraged to create challenging synthetic speech datasets for low-resource languages, supporting research on linguistic bias mitigation in ADD and anti-spoofing. Third, Indic-EcoSynth provides a high-quality resource for multilingual ADD and anti-spoofing research. Upon publication, we plan to release the metric code and Indic-EcoSynth.

7 Limitations

Scope: The study covers limited models, languages, accents, evaluators, and human participants. Additionally, carbon emissions depend upon hardware, software specifications, and region-specific carbon intensity values. Therefore, during carbon tracking, the server was exclusively used and processed one model at a time. *Server details:* Ubuntu 22.04.4 LTS, One NVIDIA RTX A6000 (CUDA 12.4), Intel Xeon E5-2630 v4 @ 2.20GHz, 128 GB RAM. While GreenVoice introduces automated evaluation metrics, these metrics are not intended to replace human evaluation. Instead, they provide a scalable and cost-effective complement when large-scale human evaluation is infeasible.

Weights and scaling factor: CRA and G-Score computations depend on the choice of weights and scaling factors, which are customizable and must be initialized based on task requirements. Section 5.1 and Section 5.2 demonstrate the impact of different scaling factors and weights on results.

Evaluator Bias: ADD and SV evaluator models may be biased towards specific languages (or accents) and generation models. We mitigate this using multiple evaluators and max-normalization.

Use multiple evaluators: Averaging the EERs of multiple evaluators helps mitigate evaluator biases. We used two ADD and two SV models.

Max-normalization: Suppose all evaluators are biased towards a particular language (or accent) L. To address this issue, GreenVoice includes a max-normalization step. It provides relative scores for generation models and helps identify the models that are most challenging for evaluators. Thus, even if evaluators show elevated EERs for L, this bias will affect all synthetic speech in L. When comparing synthetic speech in L generated by T models, evaluators will exhibit elevated EERs for each model. Since the language (or accent) L is fixed across models, max-normalized scores should still help assess their relative performance.

While bias mitigation in ADD and SV evaluators is a vital research problem, developing robust evaluators in turn requires high-quality and challenging datasets. GreenVoice enables systematic evaluation of generation models to support large-scale creation of multilingual synthetic speech datasets.

8 Ethical Considerations

Synthetic speech datasets are vital for advancing ADD and anti-spoofing research. However, such datasets may inadvertently be exploited by malicious users for refining audio deepfake generation technologies. Hence, safeguards are required to ensure ethical usage. Therefore, we plan to release Indic-EcoSynth under the CC BY-NC 4.0 license, restricting commercial use. Furthermore, GreenVoice is designed to assist responsible AI researchers in identifying generation models that pose elevated risks to ADD and SV systems. However, the framework may inadvertently be misused for unintended purposes. Nevertheless, identifying high-risk generation models enables researchers to construct more challenging benchmarks and develop more robust ADD and anti-spoofing systems. Therefore, we plan to release the GreenVoice code under the AI Pubs Research-use RAIL-S license, restricting usage to academic or research purposes. We strongly encourage responsible use of these tools and resources to advance multilingual ADD and anti-spoofing research.

[12]The supplementary material includes samples from the Indic-EcoSynth dataset.

References

Mahmoud Ahmed Mahmoud Ahmed, Karim Ahmed Elghamrawy, and Zeinab Abd El Haliem Taha. (voick): Enhancing accessibility in audiobooks through voice cloning technology. *2024 6th International Conference on Computing and Informatics (ICCI)*, pp. 46–52, 2024. URL https://api.semanticscholar.org/CorpusID:268877882.

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*, WWW '23, pp. 2033–2044, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583222. URL https://doi.org/10.1145/3543507.3583222.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pp. 2709–2720. PMLR, 2022.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model, 2024. URL https://arxiv.org/abs/2406.04904.

Ioan-Paul Ciobanu, Andrei-Iulian Hîji, Nicolae Catalin Ristea, Paul Irofti, Cristian Rusu, and Radu Tudor Ionescu. XMAD-bench: Cross-domain multilingual audio deepfake benchmark. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3109–3120, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.162. URL https://aclanthology.org/2026.findings-eacl.162/.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. Generalization ability of mos prediction networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2022. doi: 10.1109/ICASSP43922.2022.9746395.

Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024. URL https://doi.org/10.5281/zenodo.11171501.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Constance Douwes, Philippe Esling, and Jean-Pierre Briot. Energy consumption of deep generative audio models, 2021. URL https://arxiv.org/abs/2107.02621.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024. URL https://arxiv.org/abs/2412.10117.

Konstantin Gasenzer and Moritz Wolter. Towards generalizing deep-audio fake detection networks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=RGewtLtvHz.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Maurya. Benchmarking hindi-to-english direct speech-to-speech translation with synthetic data. *Language Resources and Evaluation*, 59:2613–2651, 04 2025. doi: 10.1007/s10579-025-09827-2.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9985–9998, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.493. URL https://aclanthology.org/2025.acl-long.493/.

Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 359–376, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658911. URL https://doi.org/10.1145/3630106.3658911.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh Khapra. Indicsuperb: A speech processing universal performance benchmark for indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:12942–12950, 06 2023. doi: 10.1609/aaai.v37i11.26521.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6981, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.346. URL https://aclanthology.org/2025.acl-long.346/.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4485–4495, Red Hook, NY, USA, 2018. Curran Associates Inc.

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 *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371, 2022. doi: 10.1109/ICASSP43922.2022.9747766.

Shareef Babu Kalluri et al. NISP: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6953–6957. IEEE, 2021.

Piotr Kawa, Marcin Plata, and Piotr Syga. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. In *Interspeech*, 2022. URL https://api.semanticscholar.org/CorpusID:250088985.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021. URL https://arxiv.org/abs/2106.06103.

Pavel Korshunov and Sébastien Marcel. Improving generalization of deepfake detection with data farming and few-shot learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):386–397, 2022. doi: 10.1109/TBIOM.2022.3143404.

Abhay Kumar, Kunal Verma, and Omkar More. Indiefake dataset: A benchmark dataset for audio deepfake detection. *ArXiv*, abs/2506.19014, 2025. URL https://api.semanticscholar.org/CorpusID:279999922.

Jingyi Li, Weiping Tu, and Li Xiao. FreeVC: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yuang Li, Min Zhang, Mengxin Ren, Xiaosong Qiao, Miaomiao Ma, Daimeng Wei, and Hao Yang. Cross-domain audio deepfake detection: Dataset and analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4977–4983, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.286. URL https://aclanthology.org/2024.emnlp-main.286/.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024. URL https://arxiv.org/abs/2411.01156.

Songting Liu. Zero-shot voice conversion with diffusion transformers, 2024. URL https://arxiv.org/abs/2411.09943.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023. doi: 10.1109/TASLP.2023.3285283.

Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. Speechlmscore: Evaluating speech generation using speech language model. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. URL https://api.semanticscholar.org/CorpusID:254535560.

Metavoiceio. Metavoice Source Code. https://github.com/metavoiceio/metavoice-src, 2024. Accessed: 2024-09-12.

Sheza Munir, Wassay Sajjad, Mukeet Raza, Emaan Abbas, Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14470–14480, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.861. URL https://aclanthology.org/2024.findings-acl.861/.

Nicolas M. Müller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Böttinger. Harder or Different? Understanding Generalization of Audio Deepfake Detection. In *Interspeech 2024*, pp. 2705–2709, 2024a. doi: 10.21437/Interspeech.2024-247.

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 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2024b. doi: 10.1109/IJCNN60899.2024.10650962.

Binh Nguyen, Shuju Shi, Ryan Ofman, and Thai Le. What you read isn't what you hear: Linguistic sensitivity in deepfake speech detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15741–15755, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.794. URL https://aclanthology.org/2025.emnlp-main.794/.

Riccardo Passoni, Francesca Ronchini, Luca Comanducci, Romain Serizel, and Fabio Antonacci. Diffused responsibility: Analyzing the energy consumption of generative text-to-audio diffusion models, 2025. URL https://arxiv.org/abs/2505.07615.

Victoria Popa, Andrea Morelli, Christian Di Maio, Luca Dini, Cristian Cosci, and Emanuele Fulvio Perri. Cantastoria: Enhancing audiobook engagement through adaptive soundtracks and voice cloning. In I-Hsien Ting, Panagiotis Karampelas, Min-Yuh Day, and Reda Alhajj (eds.), *Advances in Social Networks Analysis and Mining*, pp. 215–220, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-031-85240-4.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei. Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=8c50f-DoWAu.

Ravindra M. Purohit, Arth J. Shah, and Hemant A. Patil. Ggmddc: An audio deepfake detection multilingual dataset. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–6, 2024. doi: 10.1109/APSIPAASC63619.2025.10848580.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning, 2024. URL https://arxiv.org/abs/2312.01479.

Mouna Rabhi, Spiridon Bakiras, and Roberto Di Pietro. Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Syst. Appl.*, 250(C), September 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123941. URL https://doi.org/10.1016/j.eswa.2024.123941.

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. Speechbrain: A general-purpose speech toolkit, 2021. URL https://arxiv.org/abs/2106.04624.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, Singapore, 2022. IEEE. doi: 10.1109/ICASSP43922.2022.9746108.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-Aware Automatic Evaluation of Speech Generation Leveraging NLP Evaluation Metrics. In *Interspeech 2024*, pp. 4943–4947, 2024. doi: 10.21437/Interspeech.2024-1508.

Elizabeth Salesky, Julian Mäder, and Severin Klinger. Assessing evaluation metrics for speech-to-speech translation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 733–740, 2021. doi: 10.1109/ASRU51503.2021.9688073.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. IndicSynth: A large-scale multilingual synthetic speech dataset for low-resource Indian languages. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22037–22060, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1070/.

Swati Sharma, Divya V. Sharma, and Anubha Gupta. Task-lens: Cross-task utility based speech dataset profiling for low-resource indian languages, 2026. URL https://arxiv.org/abs/2602.23388.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355/.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373, 2021. doi: 10.1109/ICASSP39728.2021.9414234.

Milad Taleby Ahvanooey, Wojciech Mazurczyk, and Dongwon Lee. Socioeconomic threats of deepfakes and the role of cyber-wellness education in defense. *Commun. ACM*, 68(9):70–79, August 2025. ISSN 0001-0782. doi: 10.1145/3715317. URL https://doi.org/10.1145/3715317.

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. BharatBBQ: A multilingual bias benchmark for question answering in the Indian context. *Transactions of the Association for Computational Linguistics*, 13:1672–1692, 2025. doi: 10.1162/tacl.a.55. URL https://aclanthology.org/2025.tacl-1.75/.

Sathvik Udupa, Jesuraja Bandekar, Abhayjeet Singh, Deekshitha G, Saurabh Kumar, Sandhya Badiger, Amala Nagireddi, Roopa R, Prasanta Kumar Ghosh, Hema A. Murthy, Pranaw Kumar, Keiichi Tokuda, Mark Hasegawa-Johnson, and Philipp Olbrich. Limmits'24: Multi-speaker, multi-lingual indic tts with voice cloning. *IEEE Open Journal of Signal Processing*, 6:293–302, 2025. doi: 10.1109/OJSP.2025.3531782.

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

Praveen Srinivasa Varadhan, amogh gulati, Ashwin Sankar, Srija Anand, Anirudh Gupta, Anirudh Mukherjee, Shiva Kumar Marepally, Ankur Bhatia, Saloni Jaju, Suvrat Bhooshan, and Mitesh M Khapra. Rethinking MUSHRA: Addressing modern challenges in text-to-speech evaluation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=oYmRiWCQ1W.

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. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026, 2020. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2019.101026. URL https://www.sciencedirect.com/science/article/pii/S0885230819302700.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Voiceassistant-eval: Benchmarking ai assistants across listening, speaking, and viewing, 2025. URL https://arxiv.org/abs/2509.22651.

Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha jin Yu, Nicholas W. D. Evans, and Tomi H. Kinnunen. Sasv 2022: The first spoofing-aware speaker verification challenge. In *Interspeech*, 2022. URL https://api.semanticscholar.org/CorpusID:247762170.

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. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220, 2022. doi: 10.1109/ICASSP43922.2022.9746939.

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. Add 2023: the second audio deepfake detection challenge. *ArXiv*, abs/2305.13774, 2023. URL https://api.semanticscholar.org/CorpusID:258841572.

Jiangyan Yi, Chu Yuan Zhang, Jianhua Tao, Chenglong Wang, Xinrui Yan, Yong Ren, Hao Gu, and Junzuo Zhou. Add 2023: Towards audio deepfake detection and analysis in the wild, 2024. URL https://arxiv.org/abs/2408.04967.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Yi Zhu, Surya Koppisetti, Trang Tran, and Gaurav Bharaj. Slim: style-linguistics mismatch model for generalized audio deepfake detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

A Statistics of Indic-EcoSynth

Table 6 illustrates the per-language statistics of Indic-EcoSynth. As shown in Figure 3, the dataset has about 1,147 hours and 984 hours of male and female voice samples, respectively.

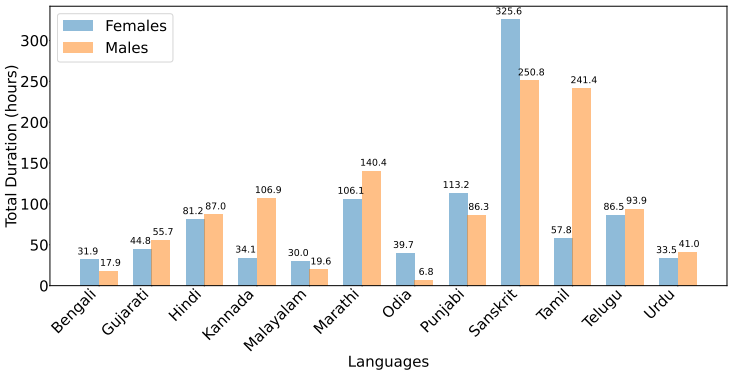

Figure 3: Per-language total duration (hours) of male and female voice samples in Indic-EcoSynth

Table 6: Statistics of Indic-EcoSynth, reporting the number of female and male speakers, total number of synthetic clips, and total speech duration (in hours) for each language.

Language	#Female	#Males	#Clips	Duration (hrs)
Bengali	18	10	27,336	49.73
Gujarati	25	34	59,660	100.44
Hindi	53	48	104,736	168.12
Kannada	16	43	59,412	141.06
Malayalam	10	7	17,094	49.58
Marathi	51	72	130,150	246.48
Odia	22	4	26,184	46.55
Punjabi	67	55	126,110	199.46
Sanskrit	100	85	192,134	576.41
Tamil	32	106	144,036	299.20
Telugu	41	43	85,728	180.38
Urdu	21	26	47,804	74.47

B Linguistic Authenticity of Odia

The accuracy values for the Odia test sets are missing in Table 5 because the VoxLingua107 ECAPA-TDNN spoken language identification model does not currently support Odia. However, the model was trained on 107 languages. Therefore, we anticipate that the language identification embeddings obtained from this model can still effectively capture the linguistic traits of Odia (Sharma et al., 2025). To qualitatively assess

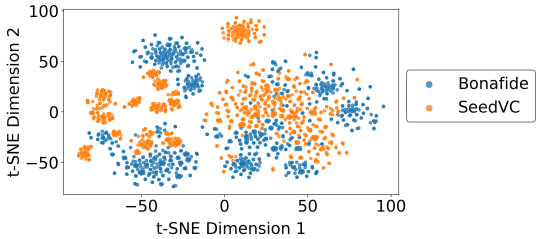

Figure 4: t-SNE visualization of bonafide and Seed-VC generated Odia speech.

this, we visualize the embeddings using t-SNE, as shown in Figure 4. We randomly selected 1,000 bonafide and 1,000 synthetic samples from the Odia language identification test set and extracted 256-dimensional language identification embeddings for each sample. We then generated a t-SNE visualization with perplexity set to 40 (Munir et al., 2024; Sharma et al., 2025).

Observation: The language identification embeddings of the bonafide and synthetic samples exhibit substantial proximity. It suggests that SeedVC-generated speech captures the linguistic traits of Odia.