# Towards Green and Inclusive Speech Processing: Understanding and Responsibly Mitigating Linguistic and Accent Biases

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

BY

**V. DIVYA SHARMA**

**PHD19012**

COMPUTER SCIENCE AND ENGINEERING

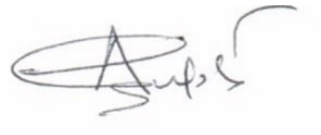
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

**April 2026**

# THESIS CERTIFICATE

This is to certify that the thesis titled **Towards Green and Inclusive Speech Processing: Understanding and Responsibly Mitigating Linguistic and Accent Biases**, submitted by **V. Divya Sharma**, to the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



**Prof. Anubha Gupta**
Thesis Supervisor
Professor
Dept. of Computer Science and Dept.
of Electronics and Communication
IIIT-Delhi, 110020

Place: New Delhi

Date: 27th April 2026

# DECLARATION

This is to be certified that the dissertation entitled **Towards Green and Inclusive Speech Processing: Understanding and Responsibly Mitigating Linguistic and Accent Biases** being submitted by **V. Divya Sharma** to the **Indraprastha Institute of Information Technology-Delhi**, for the award of degree of **Doctor of Philosophy**, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Prof. Anubha Gupta**. The study pertains to this dissertation has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

**V. Divya Sharma**
PhD19012
PhD Student
Dept. of Computer Science
IIIT-Delhi, 110020

Place: New Delhi

Date: 27th April 2026

# ACKNOWLEDGEMENTS

I am immensely grateful to my supervisor, **Prof. Anubha Gupta**, for her unwavering belief in my potential, especially during the most challenging phases of my PhD journey. Her guidance, mentorship, and constant support have profoundly shaped my academic and personal growth. At a time marked by uncertainty and stress, her encouragement became a turning point that enabled me to regain direction and continue progressing towards completing my PhD with achievements and recognitions I had never imagined possible. Beyond research, I will always cherish our daily conversations over tea, numerous celebratory treats, and the memorable trip to Vienna.

I often wondered whether I would be able to make her proud during my PhD journey. By God's grace, the day we received the Outstanding Paper Award at ACL 2025, I saw tears of happiness in her eyes and felt that I had finally achieved that goal. At the same time, I came to appreciate her humility even more. While she stands firmly with her students during difficult times, she chooses to give them full credit for their achievements. Despite my repeated requests, she declined to have an individual photograph taken with the award certificate. Instead, she celebrated the moment by expressing her appreciation and encouraging me further. I feel deeply grateful to have had the opportunity to learn from her, and I carry forward the values of resilience, compassion, and academic excellence that she embodies.

I sincerely express my deep gratitude to **Prof. Pankaj Jalote** (Founding Director, IIIT-Delhi, 2008–2018) for his valuable guidance and constructive suggestions during critical phases of my PhD. I first had the opportunity to interact with him in December 2021, when I was assigned a teaching assistantship for his course. At that time, during the third year of my PhD, I was experiencing significant stress due to publication challenges. While assisting him in preparing lecture materials, I was deeply inspired by the effort and rigor he invested in course design. Our weekly discussions, marked by his attention to detail and clarity of thought, encouraged me to approach my research with greater focus and patience, without being overly concerned about immediate outcomes.

During this period, he also kindly approved my request for leave to focus on my NAACL 2022 submission, which later became my first accepted paper and a major source of motivation. The experience of assisting him in preparing lecture slides significantly improved my ability to structure and present my research in a clear and concise manner, particularly during my time-constrained yearly reviews. His emphasis on quality over quantity, as reflected in his blog posts, significantly influenced my research philosophy. I consciously chose to focus on publishing high-quality papers in reputed venues, even if they were fewer in number.

Over the course of my PhD, I reached out to him on several occasions for guidance. Despite not being formally associated with my research, he was always approachable and supportive, and his insights and suggestions helped me navigate a critical phase of my PhD. I remain deeply grateful for his support and guidance, and I hope to continue seeking his advice and learning from his wisdom in the future.

I am deeply grateful to the ACL community for the **ACL Year-Round Mentorship program**, which played a pivotal role in my PhD journey. These monthly mentorship sessions, conducted by domain experts, promote equitable access to career guidance, support early-stage researchers, and provide broader perspectives on research in NLP. Several sessions were particularly impactful for me, especially "*When life happens: How to navigate challenging times in research?*". Other insightful sessions included topics on choosing NLP projects, navigating conferences, conducting research on a budget, effective communication, building datasets, identifying research directions, working with low-resource languages, literature review practices, and research ethics. I am deeply thankful for the guidance provided by the ACL mentors in helping me navigate various challenges during my research journey. The ACL with Love Slack community also helped me stay connected with the broader research community.

I sincerely thank my Ph.D. thesis examiners, **Prof. Dagmar Gromann** (University of Vienna), **Prof. Tanu Mitra** (University of Washington), and **Prof. Animesh Mukherjee** (Indian Institute of Technology, Kharagpur), for their valuable time and constructive feedback on our work. I also thank **Prof. Manorma Sharma** (Dean Academic, RKGITM) and **Prof. Rakesh Goel** (Director, RKGITM) for their valuable insights and suggestions during a critical phase of my PhD journey.

I am grateful to **Prof. Syamantak Das** (CSE Program Coordinator), **Mr. Raju**

**Biswas** (Admin PhD), **Mr. Ashutosh Brahma** (former Administrative Officer, Academics), **Mrs. Priti Patwal** (former Admin, CSE), and **Mr. Bhawani Shah** (Deputy Technical Officer) for their support. I sincerely thank **Prof. Arani Bhattacharya**, who was my course instructor during the Winter 2025 teaching assistantship. He kindly granted me a 45-day leave to focus on my ACL 2025 paper. I am grateful to **Prof. Dheeraj Sanghi** (former Dean of Academic Affairs, IIIT-Delhi), who interviewed me in 2016 for admission to the M.Tech program at IIIT Delhi. I also thank **IIIT-Delhi** for the research opportunity.

I convey my heartfelt thanks to my undergraduate collaborators, **Vijval Ekbote** and **Swati Sharma**. Working with them was truly a delight. I also thank my friends, **Nidhi Goyal**, **Shivani Kumar**, **Avinash Tulasi**, **Khushboo Chitre**, **Jasmeet Kaur**, and **Monika Jain**, for their support and help on various occasions. Special thanks to **Aakriti Attri**, **Agrata Singh**, **Vaibhav Sharma**, **Vardhana Sharma**, **Sarthak Kandpal**, **Ashutosh Vaish**, and members of the **SBILab** for providing a warm, respectful, and supportive environment conducive to research.

I am deeply grateful to my late maternal grandfather, **Mr. K. Lakshmanan**, for always emphasizing the importance of education and teaching me how to overcome life's adversities with resilience. His blessings continue to give me the strength to overcome the challenges that come my way. I also express my gratitude to my late paternal grandfather, **Mr. Venkatachalam Mahadevan Sharma**, whose blessings have been instrumental in enabling me to accomplish my research work. I deeply admire him for his strong will, honesty, and determination. I sincerely thank my cousin sister, **Dr. Mallika Ramakrishnan**, for her motivation and experience-based guidance whenever I was stuck during my PhD journey. I sincerely thank my brother, **Mr. V. Shrikant Sharma**, and his family for their constant encouragement and support.

Most importantly, I am indebted to my parents for their innumerable sacrifices, unwavering support, and constant motivation, which made this thesis possible. My mother, **Mrs. V. Lakshmi**, has been a constant source of strength during adversity. I deeply admire her resilience and patience in navigating life's challenges. She instilled in me the same courage, continually encouraged me to remain focused on my research, and never allowed me to give up. Words cannot fully express my gratitude to her. My father, **Mr. M. V. Sharma**, endured numerous hardships to ensure my safety and com-

fort. He has always prioritized my education above everything else. During periods of intense deadlines and sleepless nights, he would personally drive me to campus, often at odd hours despite his own fatigue, ensuring that I could focus on my work without worry. Coincidentally, the welcome reception at ACL 2025 was held on my father's birthday. I had the privilege of celebrating my first international award at ACL 2025 with my parents in Vienna, a moment made even more special when I saw the tears of pride and happiness in their eyes. My parents left no stone unturned in their efforts to ensure that I could fully focus on my research.

This journey has strengthened my belief in divine guidance, as reflected in the *Bhagavad Gita*, that balance is restored when imbalance arises. I am deeply devoted to Kanha, who is my eternal source of guidance and strength in navigating life's challenges. I am also deeply indebted to the divine mother Goddess Shakti, who is my guiding force and source of strength, protection, and resilience through life's difficulties. I remain deeply grateful to **God** for guidance, protection, and strength throughout this PhD journey, which I believe has been shaped by divine grace.

*I dedicate this thesis to my parents, whose support and sacrifices made this work possible.*

*V. Divya Sharma*

2nd May 2026

# ABSTRACT

KEYWORDS: Inclusive AI ; Green AI ; Bias Mitigation ; Responsible AI ; Synthetic Speech

High-quality synthetic speech has transformative potential for accessibility, education, entertainment, and personalized human–computer interaction. However, it also poses serious risks: synthetic voices can be exploited for audio deepfakes and impersonation attacks. These risks are magnified in multilingual and low-resource settings, where audio deepfake detection (ADD) and speaker verification (SV) systems exhibit pronounced linguistic biases, and the scarcity of large-scale, publicly available datasets limits the development of robust, fair, and inclusive models. Moreover, existing methods for evaluating synthetic speech quality rely primarily on human studies, which are costly, difficult to scale, and often lack reproducibility. Additionally, synthetic speech generation models incur significant carbon emissions, yet environmental sustainability remains largely overlooked. Together, these challenges highlight a critical need for datasets, evaluation frameworks, and bias-mitigation methods that can enable responsible, inclusive, and environmentally conscious speech technologies.

To address these gaps, this thesis makes the following key contributions: First, we introduce IndicSynth, a large-scale synthetic speech dataset covering 12 low-resource Indian languages to support multilingual ADD and anti-spoofing research. IndicSynth balances realistic voice mimicry and synthetic diversity. Using IndicSynth, we demonstrate the vulnerability of existing ADD and SV models against synthetic speech attacks. Human evaluation further validates the dataset quality, underscoring the dataset's utility for security-focused applications. Second, we present Task-Lens, a cross-task profiling framework to mitigate task-resource gaps for underrepresented languages. Using Task-Lens, we profile 34 Indian speech datasets, including IndicSynth, covering 26 languages and eight downstream tasks, based on available metadata. Third, we propose FAtNet and EcoSpeak, which are cost-efficient methods for mitigating linguistic biases in speaker verification, addressing fully and partially cross-lingual scenarios while

incorporating Green AI principles by reporting carbon emissions. Finally, we introduce GreenVoice, an automated environment-aware evaluation framework for synthetic speech generation models. GreenVoice cost-effectively highlights high-performing and sustainable generation models for large-scale synthetic speech dataset creation, thus enabling multilingual ADD and anti-spoofing research across more underrepresented languages and accents, beyond IndicSynth. Together, these contributions provide the foundations for building and evaluating speech technologies that are robust, equitable, and inclusive across languages and accents, while promoting environmentally responsible practices and supporting their reliable use in real-world applications.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# Introduction

## 1.1 Motivation

Speech technologies have become increasingly pervasive, shaping how humans interact with machines, access information, and engage with digital content. In particular, synthetic speech produced by text-to-speech (TTS) and voice conversion (VC) models has seen remarkable advances in recent years, yielding highly realistic, human-like voices [8]. These technologies hold transformative potential across a range of applications, including accessibility for individuals with speech disorders, educational tools, audiobooks, personalized virtual assistants, entertainment, and language preservation [15, 133, 193].

The rise of synthetic speech also brings significant challenges. High-fidelity synthetic voices can be exploited for malicious purposes, including impersonation attacks, audio deepfakes, and the spread of misinformation [112, 70, 78, 108, 47, 133]. In addition, state-of-the-art (SOTA) systems for audio deepfake detection (ADD) and speaker verification (SV) are often optimized for high-resource languages, leading to poor generalization for low-resource and underrepresented languages [8, 185, 7, 180]. Similarly, voice cloning models frequently exhibit linguistic and accent biases, resulting in uneven performance across different speakers and populations [201, 184]. These challenges are further compounded by the substantial computational resources required to train and deploy modern speech models, raising concerns about sustainability and environmental impact [158, 184, 187]. As summarized in Figure 1.1, these applications and challenges illustrate the motivation behind the present research.

This thesis investigates challenges in multilingual and low-resource speech processing, addressing issues of inclusivity, fairness, and sustainability across a range of speech technologies. Specifically, it focuses on (i) creating and evaluating multilingual synthetic speech datasets to support audio deepfake detection and anti-spoofing tasks; (ii) mitigating cross-lingual biases in speaker verification systems; (iii) investigating accent

biases in voice cloning models; and (iv) developing environmentally responsible frameworks and evaluation metrics for benchmarking synthetic speech models. Collectively, these efforts advance inclusive, responsible, and sustainable approaches to speech processing, addressing both data scarcity and model-level limitations in low-resource settings.

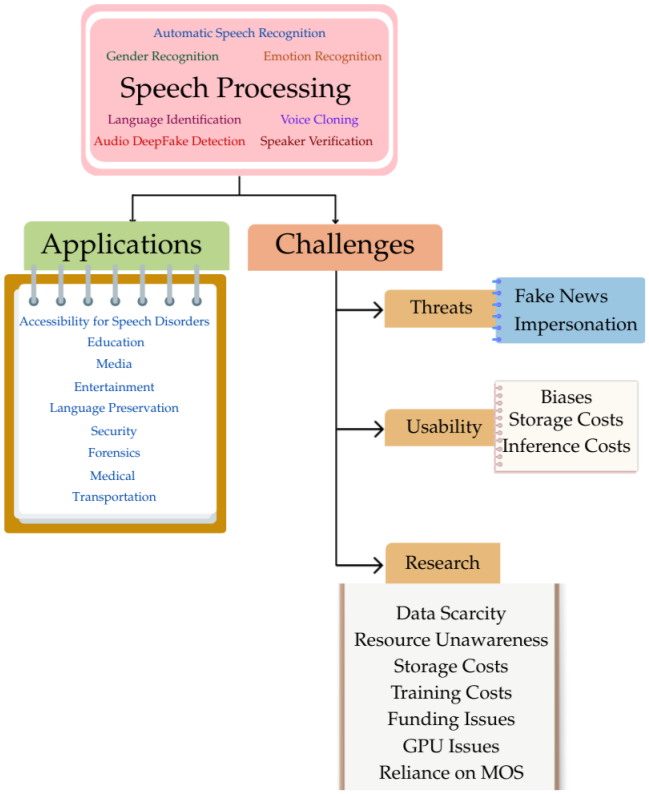



Figure 1.1: Overview of speech technologies, their applications, and associated challenges

## 1.2 Background

This thesis investigates speech processing in multilingual and low-resource settings, with a focus on security, authentication, and the opportunities and challenges introduced by synthetic speech. To provide context, we briefly define key terms and technologies foundational to this work, including text-to-speech (TTS), voice conversion (VC), speaker verification (SV), audio deepfake, audio spoof, and audio deepfake detection (ADD).

### 1.2.1 Text-to-Speech

TTS systems generate synthetic speech from a text transcript while conditioning on a target speaker's voice sample. The generated speech articulates the given text while attempting to mimic the voice of the target speaker.

### 1.2.2 Voice Conversion

VC systems take a source speech sample and a target speaker sample as input. The system produces a synthetic speech sample that preserves the content of the source input while mimicking the target speaker's voice.

### 1.2.3 Speaker Verification

SV systems determine whether two speech recordings belong to the same speaker. SV is critical for authentication, security, and anti-spoofing applications. Modern SV systems often rely on deep Convolutional Neural Networks (CNNs) to model speaker-specific characteristics.

### 1.2.4 Audio Deepfakes

Audio deepfakes are realistic synthetic speech samples generated to deceive humans and spread misinformation or fake news that can cause public unrest or panic [112].

### 1.2.5 Audio Spoofs

An audio spoof is a synthetic speech recording that closely mimics the voice of a particular target speaker. Such recordings are misused to deceive speaker verification systems, leading to impersonation and identity theft [78, 112].

### 1.2.6 Audio Deepfake Detection

ADD systems aim to distinguish between bonafide (genuine) and synthetic speech. ADD is a growing research area driven by security concerns, as high-quality synthetic speech increasingly poses a threat to authentication systems, trust in media, and public discourse [112, 78, 8, 185, 7, 180].

## 1.3 Challenges in Speech Processing

Despite advances in speech technologies, several persistent challenges limit their inclusivity, fairness, and sustainability.

### 1.3.1 Inclusivity challenges

Most state-of-the-art speech processing models, including security-sensitive systems such as speaker verification (SV) and audio deepfake detection (ADD), are trained predominantly on high-resource languages. This limited coverage results in poor performance for low-resource and underrepresented languages, as well as for speakers with diverse accents or dialects [108, 180, 201]. The lack of inclusivity not only reduces generalization to unseen populations but also introduces biases in security-critical applications. For instance, SV systems may incorrectly reject genuine speakers from underrepresented groups or accept impostors exploiting cross-lingual mismatches. Similarly, ADD models trained on a narrow set of languages struggle to detect synthetic speech in other languages, increasing the risk of malicious misuse. Addressing these inclusivity and bias challenges is critical for building robust, fair, and trustworthy speech technologies [186].

### 1.3.2 Data Scarcity and Task-Specific Limitations

The absence of large scale, multilingual, and diverse datasets hinders research and system development. Low-resource languages, particularly in India, have limited publicly available data, which affects SV system robustness, and ADD performance. Beyond language coverage, the lack of awareness of existing resources and scarcity of resources

for specific tasks, such as speech emotion recognition, automatic speech recognition, or audio deepfake detection, limits research progress and the development of reliable systems [85].

### 1.3.3 Sustainability and Green AI Considerations

Large-scale speech models often comprise millions or even billions of parameters and require substantial computational power for training and inference. These demands result in significant energy consumption and carbon emissions. The NLP community has highlighted that training a single large model, such as BERT, can produce carbon emissions equivalent to a trans-American flight [158]. Similar analyses are largely missing in speech processing research, despite the increasing size and complexity of modern speech models. Despite growing awareness of Green AI principles, environmental sustainability has received limited attention in the field.

## 1.4 Linguistic Bias

Existing audio deepfake detection and speaker verification models are often affected by domain-specific biases. As a result, models trained on one domain tend to perform poorly on unseen or out-of-domain data. Domain-specific biases are not only limited to audio deepfake detection and speaker verification, but have also been observed across other speech processing tasks. Speech processing has diverse real-world applications. Therefore, addressing these biases is important, especially in security-sensitive tasks like audio deepfake detection and speaker verification.

Linguistic bias is a form of domain-specific bias, where models trained on one language exhibit performance degradation when evaluated on unseen languages. With about 7,000 languages spoken worldwide, mitigating linguistic bias is crucial for improving the global usability of speech technologies. This is especially important in linguistically diverse countries like India. India has 22 scheduled language, 121 major languages, and 1369 mother tongues spoken by over 1.4 billion people [51, 167].

## 1.5 Pillars of this Thesis

To address the challenges above, this thesis is organized around four key pillars, spanning dataset creation, model development, training strategies, and sustainable evaluation, as shown in Figure 1.2.

### 1.5.1 Datasets

- **IndicSynth:** We introduce IndicSynth, a large-scale multilingual synthetic speech dataset covering 12 low-resource Indian languages, consisting of 4,000 hours of speech from 989 target speakers. IndicSynth includes mimicry and diversity subsets to balance realistic voice imitation with broad diversity, thus facilitating multlingual audio deepfake detection and anti-spoofing research.
- **Task-Lens:** We propose Task-Lens, a cross-task profiling framework to evaluate the applicability of speech datasets across multiple downstream tasks, enabling resource-driven prioritization for underrepresented languages.

### 1.5.2 Model Design

- **FAtNet:** We propose FAtNet, a cost-efficient approach designed to mitigate fully cross-lingual biases in SV by leveraging lightweight frame-level embeddings and attention mechanisms. Fully cross-lingual biases can occur when both recordings in the SV trial pair are in an unseen target language that is different from the training language.
- **EcoSpeak:** Extends bias mitigation to partially cross-lingual scenarios using contrastive linguistic attention and bias-correcting mechanisms, providing environmentally aware SV solutions. Partially cross-lingual bias can occur when one trial pair recording is in the training language, whereas the other is in an unseen target language.

### 1.5.3 Training Strategies

- Investigate how training on strongly related but limited datasets versus weakly related but diverse datasets affects generalization.
- Develop training data balancing strategies to reduce linguistic biases in SV models, improving fairness and robustness across languages and accents.

### 1.5.4 Environmentally Responsible Evaluation

**GreenVoice:** We propose GreenVoice, a novel automated evaluation framework that evaluates synthetic speech generation models through a composite metric, the $G$-Score, integrating cloning quality (realism and similarity) with environmental impact (carbon emissions). GreenVoice supports the selection of high-performing and sustainable models for large-scale audio deepfake detection and anti-spoofing dataset creation.

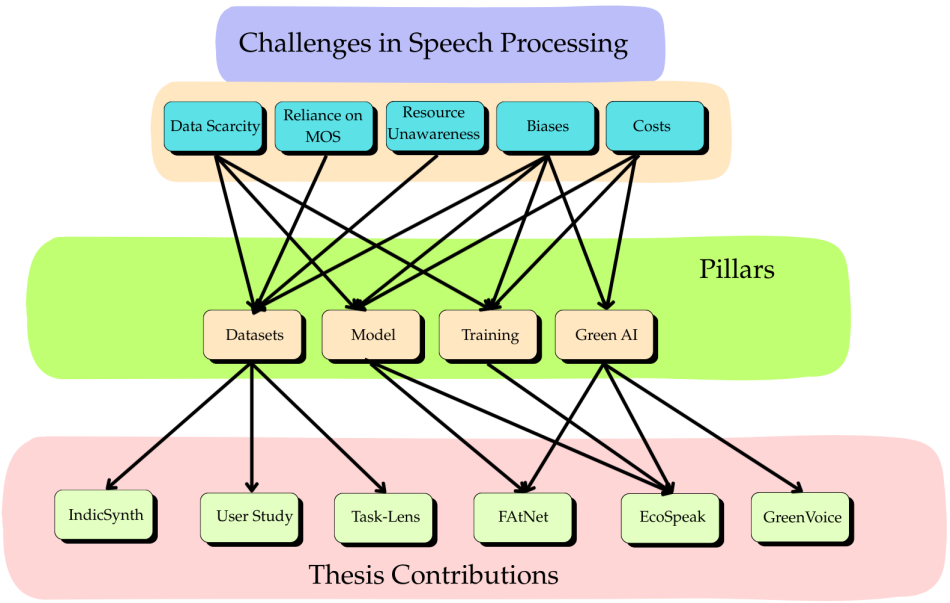



Figure 1.2: Overview of Thesis Contributions

### 1.5.5 Relation between the Pillars of the Thesis

IndicSynth provides a large-scale dataset to support multilingual audio deepfake detection (ADD) and anti-spoofing research in Indian languages. Task-Lens extends this investigation by exploring the cross-task utility of IndicSynth and several other Indian speech resources for underrepresented tasks based on the available metadata in these datasets. While experiments on IndicSynth demonstrate the vulnerability of existing ADD and speaker verification models against multilingual synthetic speech attacks, FAtNet and EcoSpeak explore solutions for cost-effectively mitigating linguistic biases in speaker verification. EcoSpeak further discusses the concept of Green speech processing in regard to speaker verification. GreenVoice extends this discussion with respect to selecting sustainable generation models for large-scale synthetic speech dataset creation. Overall, GreenVoice addresses the common challenges faced by dataset cre-

ators related to human evaluation of synthetic speech, financial costs, resource constraints, and environmental impact. Thus, enhancing research inclusivity and promoting the sustainable development of large-scale multilingual synthetic speech datasets for ADD and anti-spoofing research.

## 1.6 Research Objectives

Based on the challenges and pillars, the main objectives of this thesis are:

1. Develop a multilingual synthetic speech dataset to facilitate multilingual ADD and anti-spoofing research for low-resource Indian languages.
2. Conduct human-centered studies to assess perceptual naturalness, mimicry quality, and detectability of synthetic speech.
3. Mitigate task-resource gaps in underrepresented Indian languages.
4. Mitigate linguistic biases in speaker verification systems in a cost-efficient and environmentally responsible manner.
5. Propose automated, scalable, and environmentally aware evaluation framework for synthetic speech generation models.
6. Provide practical guidance on overcoming common challenges in large-scale ADD and anti-spoofing dataset creation.

## 1.7 Organization of the Thesis

The thesis is organized as follows:

- **Chapter 2:** Introduces IndicSynth, highlighting data scarcity in low-resource Indian languages and the impact of linguistic biases on ADD and anti-spoofing systems.
- **Chapter 3:** Presents human perception studies to evaluate the naturalness and mimicry quality of synthetic speech in IndicSynth.
- **Chapter 4:** Introduces Task-Lens, a cross-task profiling framework to assess dataset utility across multiple speech tasks.
- **Chapter 5:** Proposes FAtNet, a cost-efficient framework to mitigate fully cross-lingual biases in speaker verification.
- **Chapter 6:** Extends bias mitigation to partially cross-lingual scenarios via EcoSpeak, incorporating Green AI principles.

- **Chapter 7:** Presents GreenVoice, an automated environment-aware evaluation framework for synthetic speech generation models, integrating quality and sustainability for large-scale ADD and anti-spoofing dataset creation.
- **Chapter 8:** Concludes with overarching findings, limitations, and future research directions.

## 1.8 Publications

The research presented in this thesis is based on the following peer-reviewed publications:

1. **Divya V. Sharma**, Vijval Ekbote, Anubha Gupta. 2025. IndicSynth: A Large-Scale Multilingual Synthetic Speech Dataset for Low-Resource Indian Languages. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (**ACL 2025**, Volume 1: Long Papers), pages 22037–22060, Vienna, Austria. Association for Computational Linguistics. **[Outstanding Paper Award]**
2. **Divya V. 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.
3. **Divya V. Sharma** and Arun Balaji Buduru. 2022. FAtNet: 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.
4. Swati Sharma, **Divya V. Sharma**, Anubha Gupta. Task-Lens: Cross-Task Utility Based Speech Dataset Profiling for Low-Resource Indian Languages. In International Conference on Language Resources and Evaluation: **LREC2026**. **[Accepted]**.
5. **Divya V. Sharma**, Vijval Ekbote, Anubha Gupta. Quality and Sustainability Metrics for Multilingual Audio Deepfake Detection and Anti-Spoofing Dataset Creation. Work in progress.

## 1.9 Honors and Awards

The work presented in Chapter 2, centered on the IndicSynth dataset, received the **Outstanding Paper Award** at *ACL 2025*. The same work was subsequently invited for

presentation at the *ACM India ARCS 2026*, where it received the **Distinguished Poster Recognition**.

IndicSynth has also seen wide adoption by the research community, with **31,007 downloads** on HuggingFace within nine months of its release.

# CHAPTER 2

# IndicSynth: Multilingual Synthetic Speech for Deepfake Detection

Recent advances in synthetic speech generation technology has enabled the creation of highly realistic speech that closely imitates human voices. While these technologies offer exciting possibilities, they also pose significant risks, including identity theft and the dissemination of misinformation. As a result, there is an urgent need for robust and generalizable audio deepfake detection (ADD) and anti-spoofing systems. However, existing models often suffer from linguistic bias—models trained on one language typically perform poorly when tested on out-of-domain languages. This limitation reduces their effectiveness and underscores the necessity of multilingual synthetic speech datasets to support bias mitigation research. Yet, most publicly available datasets focus only on English or Chinese, leaving a critical gap for other languages. The data scarcity challenge is particularly severe in linguistically diverse regions such as India.

To address this, we present *IndicSynth*[1], a large-scale dataset comprising 4,000 hours of synthetic speech from 989 target speakers (456 female and 533 male) across 12 low-resource Indian languages. The dataset provides detailed metadata, including gender information and speaker identifiers. Experimental evaluations confirm that IndicSynth serves as a valuable resource for advancing multilingual ADD and anti-spoofing research. The dataset is publicly available at: `https://huggingface.co/datasets/vdivyasharma/IndicSynth`.

## 2.1 Introduction

Recent progress in text-to-speech (TTS) and voice conversion (VC) technologies has made it possible to generate synthetic speech that closely resembles human voices [8].

[1]This chapter presents the following paper:
Divya V Sharma, Vijval Ekbote, and Anubha Gupta. 2025. IndicSynth: A Large-Scale Multilingual Synthetic Speech Dataset for Low-Resource Indian Languages. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22037–22060, Vienna, Austria. Association for Computational Linguistics.

These synthetic recordings are increasingly used in domains such as assistive communication, media, entertainment, and education [15, 133, 193]. However, alongside these benefits lies a significant risk of misuse, including the spread of misinformation, impersonation of public figures, financial fraud, and other malicious activities [112, 70]. Since modern synthetic speech generative systems can create speech that convincingly imitates real voices, they can deceive not only human listeners but also security systems like speaker verification. Consequently, developing reliable methods to identify synthetic (fake) speech has become essential to counter such risks [112].

An audio deepfake refers to a synthetic speech recording that sounds realistic enough to mislead human listeners and can be exploited to disseminate misinformation, potentially triggering public unrest or panic [112]. In contrast, audio spoofing encompasses techniques that generate synthetic speech imitating a specific speaker's voice. Such methods are frequently abused to bypass biometric security systems, including speaker verification [78, 112], thereby enabling impersonation and identity theft, as shown in Figure 2.1. A notable case occurred in 2019, when fraudsters used spoofed audio of a corporate executive to orchestrate a financial scam worth 243,000 USD [108, 47]. In another incident, audio deepfake-driven fraud led to a loss of 35 million USD for a UAE-based company [133]. Beyond financial crimes, audio deepfakes are also weaponized to damage the reputations of public figures and to sway voter sentiment during elections [78]. These risks underscore the urgent need for effective audio deepfake detection and anti-spoofing technologies to safeguard individuals and society.


Figure 2.1: Audio deepfakes are realistic synthetic speech recordings that may be misused to deceive humans and spread misinformation, causing public unrest and panic. In contrast, audio spoofs are synthetic speech recordings that closely mimic the target speaker's voice. Audio spoofing is misused to deceive speaker verification systems, leading to impersonation and identity theft.

Developing effective audio deepfake detection (ADD) models requires access to realistic synthetic speech datasets. Yet, the majority of available datasets are concen-

trated in high-resource languages like English and Chinese [112, 8, 108, 47, 79]. As a result, much of the existing ADD research has focused on these languages [194]. However, models trained on such datasets often experience substantial drops in performance when applied to out-of-domain languages [112]. This underscores the urgent need for synthetic speech datasets in low-resource languages to improve the generalizability and worldwide applicability of ADD systems. The demand is particularly pressing in linguistically diverse countries like India, where 22 constitutionally recognized languages are spoken by over a billion people, and nearly 75% of the population encounters some form of deepfake content annually [64, 160].

In this chapter, we introduce IndicSynth, a large-scale multilingual synthetic speech dataset covering 12 low-resource Indian languages. The dataset contains around 4,000 hours of synthetic audio generated from 989 target speakers (456 female and 533 male). It is enriched with detailed metadata, including gender and speaker identifiers. To enhance its utility for multilingual audio deepfake detection (ADD) and anti-spoofing research, IndicSynth is divided into two subsets: mimicry and diversity. The mimicry subset consists of synthetic samples (audio spoofs) that closely mimic their corresponding bonafide target voices. In contrast, the diversity subset offers a broader variety of realistic synthetic voices (audio deepfakes). The dataset was generated using state-of-the-art (SOTA) text-to-speech (TTS) and voice conversion (VC) models applied to publicly available bonafide speech data. After generation, we evaluated the dataset along multiple dimensions: we examined whether SOTA ADD models could reliably classify IndicSynth audios as synthetic, tested their linguistic authenticity using a SOTA language identification model, and investigated whether IndicSynth's mimicry subset could deceive SOTA speaker verification systems through impersonation attacks. All experiments were carried out using publicly available models to ensure reproducibility.

**Summary of Chapter Contributions:**

1. This chapter introduces IndicSynth, a multilingual synthetic speech dataset for 12 low-resourced Indian languages. The dataset contains about 4,000 hours of audio from 989 target speakers (456 female and 533 male) and organizes the data into mimicry and diversity subsets.
2. The study evaluates linguistic bias in state-of-the-art audio deepfake detection (ADD) models and examines the vulnerability of state-of-the-art speaker verification systems to impersonation attacks using multilingual audio spoofs. It further investigates the utility of IndicSynth for building more generalizable ADD and anti-spoofing models.

3. This chapter evaluates the linguistic authenticity of IndicSynth both qualitatively and quantitatively using t-SNE visualizations and a state-of-the-art language identification model.

## 2.2 Related Works

### 2.2.1 Audio DeepFake Detection and Anti-Spoofing

The rise of audio deepfakes and spoofing-related fraud has motivated the research community to organize Audio DeepFake Detection (ADD) and ASVspoof challenges [192, 193, 178, 94]. Despite these efforts, a major challenge that remains underexplored is the limited generalizability of ADD and anti-spoofing models across out-of-domain scenarios [83, 195, 185, 78, 110]. Models trained on speech datasets in a single language often experience substantial drops in accuracy when evaluated on other languages. This linguistic bias significantly limits their effectiveness in real-world applications [146, 145].

### 2.2.2 Lack of Datasets

To address linguistic bias in ADD and anti-spoofing systems, researchers explore domain adaptation and data augmentation strategies [8, 185]. However, these approaches rely on the availability of synthetic speech datasets in the target languages. Most existing datasets, such as FakeAVCeleb, ASVspoof 2019, and ADD 2023, are primarily in English or Chinese [193, 112, 108, 8, 79, 178]. To diversify resources, researchers introduced the Urdu audio deepfake detection dataset with 16,830 spoofed recordings [108], and the WaveFake dataset with 196 hours of synthetic speech in English and Japanese [47]. Other efforts include the MLAAD and MLADDC datasets [112, 143], though these lack crucial metadata such as gender details and target speaker identifiers. Gender information is essential for studying demographic and gender-related biases in ADD [188, 70, 78, 54], while speaker identifiers are vital for developing defenses against impersonation attacks. The MADD dataset adds further diversity with 155.66 hours of synthetic audio across six languages [130]. To bridge the gap, we introduce **IndicSynth**, a large-scale multilingual synthetic speech dataset that provides not only

extensive coverage across low-resource Indian languages but also detailed gender information and target speaker identifiers.

## 2.3 IndicSynth: Generation and Overview

This chapter introduces IndicSynth, a novel large-scale multilingual synthetic speech dataset. IndicSynth comprises nearly 4,000 hours of synthetic speech across 12 low-resource Indian languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu, as shown in Figure 2.2. This section outlines the process used to generate the dataset along with its statistical characteristics.

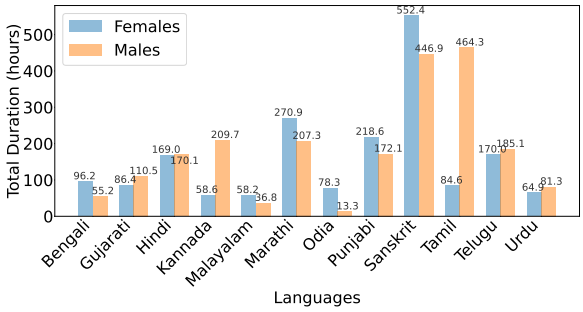



Figure 2.2: Total duration (in hours) of synthetic male and female speech in IndicSynth for each target language.

### 2.3.1 IndicSynth Generation Methodology

To create IndicSynth, we begin with the IndicSuperb dataset [64]. IndicSuperb is available under the Creative Commons CC0 ("no rights reserved") license. The dataset includes bonafide (real) speech recordings along with their corresponding transcripts for the 12 target languages. Using IndicSuperb as the foundation, we apply publicly available text-to-speech (TTS)–based voice cloning models and voice conversion (VC) models to generate synthetic speech, as shown in Figure 2.3[2]. TTS and VC models are widely employed in synthetic data generation [205]. TTS models take as input a bonafide target speech sample ($v_{\text{tgt}}$) and its transcript ($t_{\text{txt}}$), and produce a synthetic recording (${v_{\text{tgt}}}^{\text{tts}}$) as output, as defined in Equation 2.1:

$$ {v_{\text{tgt}}}^{\text{tts}} = TTS(t_{\text{txt}}, v_{\text{tgt}}) \tag{2.1} $$

[2]Models used for IndicSynth generation: `https://github.com/coqui-ai/TTS`. The complete training data for these models is not disclosed.

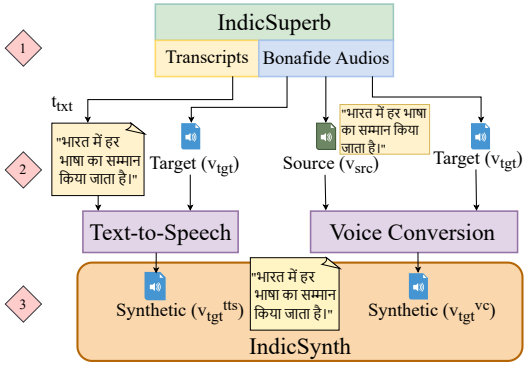



Figure 2.3: IndicSynth's generation methodology. Publicly available text-to-speech and voice conversion models were applied to the IndicSuperb dataset (licensed under CC0, 'no rights reserved') to create IndicSynth.

The synthetic speech generated by TTS models (${v_{tgt}}^{tts}$) articulates the transcript ($t_{txt}$) while attempting to mimic the target speaker's voice ($v_{tgt}$). In contrast, voice conversion (VC) models operate differently: they take as input a bonafide source speech sample ($v_{src}$) along with a bonafide target speech sample ($v_{tgt}$). The output is a synthetic recording (${v_{tgt}}^{vc}$) that resembles the target speaker's voice, as shown in Equation 2.2.

$$ {v_{tgt}}^{vc} = VC(v_{src}, v_{tgt}) \tag{2.2} $$

The synthetic speech (${v_{tgt}}^{vc}$) articulates the speech content from the source audio ($v_{src}$) while attempting to mimic the target speaker's voice ($v_{tgt}$).

### 2.3.2 Mimicry and Diversity

IndicSynth comprises two categories of synthetic data, as outlined in Table 2.1:

1. *Mimicry*: The mimicry subset contains audio spoofs that closely mimic bonafide target voices. This subset is useful for evaluating the susceptibility of speaker verification models to impersonation attacks [108].
2. *Diversity*: The diversity subset contains realistic synthetic audios with low resemblance to target voices, resulting in greater voice variation. It can be useful for training audio deepfake detection models to achieve better generalization across out-of-domain languages.

| 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 2.1: Statistical summary of IndicSynth, including generative model, subset type, counts of male and female target speakers, number of audio clips, and total synthetic audio duration (hours) for each language.

### 2.3.3 IndicSynth Generation

For the mimicry subset, synthetic data was generated by fine-tuning the XTTS-v2 model on the IndicSuperb dataset for 10 target languages: Bengali, Gujarati, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu[3]. Using the same bonafide dataset (IndicSuperb) for both fine-tuning and synthetic data generation promotes close resemblance between synthetic and target voices.

In contrast, the diversity subset was created without fine-tuning on IndicSuperb, relying instead on direct outputs from TTS and VC models. Specifically, the publicly available Coqui VITS model (trained on undisclosed Bengali data) was used to produce Bengali synthetic samples for this subset [45]. Furthermore, for all 12 target languages, synthetic data was generated with the publicly available XTTS-v2 text-to-speech model

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

and the FreeVC24 voice conversion model [45], as summarized in Table 2.1.

For IndicSynth construction, bonafide audios and transcripts were randomly sampled. Additionally, when applying voice conversion, source and target speakers were matched by gender to ensure high-quality outputs.

### 2.3.4 Metadata Details

IndicSynth includes separate metadata files for each generative model and target language. For TTS-based generation, the metadata includes the target speaker ID, ID of the bonafide reference sample, speaker gender, transcript, and ID of the generated synthetic clip. For VC-based generation, it records the source speaker ID, ID of the bonafide IndicSuperb source clip, target speaker ID, ID of the bonafide IndicSuperb target clip, speaker genders, and the ID of the resulting synthetic clip. This structured metadata also enables investigations of gender bias in multilingual audio deepfake detection (ADD) [188, 70]. More broadly, pairing bonafide data from IndicSuperb with synthetic data from IndicSynth supports research on multilingual ADD and anti-spoofing.

## 2.4 Evaluation of IndicSynth

### 2.4.1 IndicSynth for Audio DeepFake Detection

Audio deepfake detection (ADD) systems take a speech recording as input and classify it as either bonafide (real) or synthetic (fake). These systems are critical for mitigating the risks of misinformation and fake news generated through synthetic audio. Consequently, there is an urgent need for ADD models that are both inclusive and capable of generalizing to previously unseen languages. An essential step toward this goal is to benchmark state-of-the-art (SOTA) models on datasets from such languages, which can also reveal potential biases. To this end, we benchmark three publicly available SOTA ADD models—Aasist, Aasist-L, and RawNet2 [72, 161][4]—using IndicSynth. These models, originally trained on the English LA partition of the ASVspoof 2019 challenge

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

dataset [178], were evaluated on IndicSynth without any fine-tuning.

**Setup:** For each target language and generative model, distinct test sets were constructed, as shown in Table 2.2. Each set consists of 4,000 bonafide female samples, 4,000 bonafide male samples, 4,000 synthetic female samples, and 4,000 synthetic male samples, selected randomly. The bonafide samples were sourced from IndicSuperb, while the synthetic ones came from IndicSynth. These combined sets are referred to as *IndicSynth-IndicSuperb* test sets.

| | | 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 2.2: Performance of audio deepfake detection (ADD) models on IndicSynth-IndicSuperb test sets without domain adaptation. For each target language and ADD model, the maximum Equal Error Rate (EER%) across generative models is shown in bold. The results indicate that, in the absence of domain adaptation, ADD models exhibit higher EER% on these test sets. Incorporating multilingual synthetic datasets like IndicSynth for training can improve the robustness and generalizability of ADD models.

**Evaluation Metric:** The False Match Rate (FMR) and False Non-Match Rate (FNMR) are commonly used for assessing biometric systems. FMR refers to the proportion of synthetic audio samples that an ADD model incorrectly classifies as bonafide,

while FNMR represents the proportion of bonafide samples that are mistakenly classified as synthetic. Both FMR and FNMR vary depending on the decision threshold. At a specific threshold, these two error rates become equal, and this value is termed the Equal Error Rate (EER). EER is a widely accepted evaluation metric in audio deepfake detection [178, 94]. Accordingly, we benchmark ADD models using EER (%), where a lower EER indicates better discrimination between bonafide and synthetic speech.

**Observations:** Aasist and Aasist-L reported Equal Error Rates (EER) of 0.83% and 0.99% on the LA evaluation set of ASVspoof 2019 [72], while RawNet-2 achieved an EER of 22.38% in the DF track of ASVspoof 2021 [94]. In contrast, as shown in Table 2.2, these benchmark models produce much higher EERs on the IndicSynth-IndicSuperb test sets. The elevated EER values suggest that the models struggled to differentiate between bonafide and synthetic audio. In addition, Figure 2.4 presents the Receiver Operating Characteristic (ROC) curves, which reveal extremely low Area Under the Curve (AUC) scores for the Malayalam test set generated using XTTS-v2. Such low AUC values further confirm the weak discriminative capacity of the models. Overall, these findings highlight a substantial drop in performance when benchmark ADD models face unseen language test sets. Training on large-scale multilingual synthetic datasets can mitigate this issue by improving model generalization [112]. Hence, IndicSynth can serve as a valuable resource for building more robust and generalizable ADD systems.

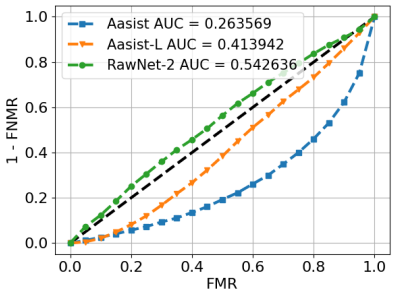



Figure 2.4: Receiver Operating Characteristic (ROC) curve for the Malayalam IndicSynth-IndicSuperb test set generated with XTTS-v2. A lower Area Under the Curve (AUC%) reflects weaker discriminative ability of the ADD models.

### 2.4.2 Linguistic Authenticity of IndicSynth

We next examine whether the synthetic speech in IndicSynth effectively preserves the linguistic characteristics of the target languages. To this end, we constructed test sets

of bonafide (IndicSuperb) and synthetic (IndicSynth) speech for each generative model and language, as shown in Table 2.3. Each test set comprises 8,000 audio clips, evenly distributed across male and female speakers. We then assess the linguistic authenticity of IndicSynth by applying the publicly available VoxLingua107 ECAPA-TDNN language identification model [170, 134] to these sets[5]. This model is trained on the VoxLingua107 dataset, which covers recordings from 107 languages [171].

| Language | Source | Accuracy | Δ 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 2.3: Language identification results. We evaluate 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.

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

**Observations:** Table 2.3 reports the accuracy of the language identification model across the test sets. Most sets achieve over 80% accuracy. We further analyze the difference between bonafide and synthetic test sets, defined as $\Delta$ Accuracy% = $\text{Accuracy}_{\text{synthetic}}$ − $\text{Accuracy}_{\text{bonafide}}$. For the majority of languages, this accuracy drop remains below 10%. Notably, Bengali synthetic speech generated by FreeVC24 and VITS attains higher accuracy than bonafide speech, suggesting that these models were trained on diverse Bengali datasets.

**Qualitative evaluation:** The VoxLingua107 ECAPA-TDNN spoken language identification model does not include Odia among its supported languages, though it has been trained on 107 languages. Consequently, its embeddings are still expected to capture relevant linguistic traits. To assess this, we extracted 256-dimensional language identification embeddings from the Odia test sets and applied t-SNE with a perplexity of 40 [108], as shown in Figure 2.5. The visualization reveals no clear distinction between bonafide (IndicSuperb) and synthetic (IndicSynth) embeddings, suggesting that the IndicSynth-Odia subset successfully reflects the linguistic characteristics of Odia.

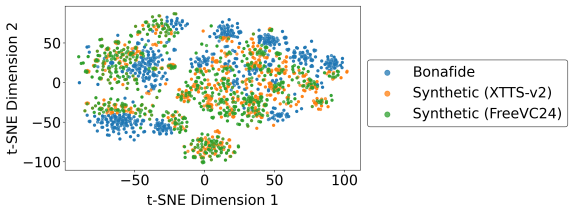



Figure 2.5: t-SNE visualization of bonafide (IndicSuperb) and synthetic (IndicSynth) Odia data. The plot suggests that the IndicSynth-Odia subset successfully preserves the linguistic characteristics of Odia.

### 2.4.3 Utility of the Mimicry Subset

Speaker verification (SV) systems take two speech recordings as input and determine whether they belong to the same speaker. Each input pair is referred to as a trial pair. These systems play an important role in domains such as forensics, business, e-commerce, and access control. However, advances in voice cloning technologies make it possible to create synthetic speech that closely imitates a target speaker's voice. Such synthetic recordings, often termed audio spoofs, can be exploited to bypass SV systems and carry out impersonation attacks. To counter this, fine-tuning SV models on multilingual synthetic speech datasets can improve their robustness and generalizability against previously unseen spoofs. Motivated by this hypothesis, our experiment

investigates the usefulness of IndicSynth's mimicry subset in strengthening SV models. Specifically, we test whether synthetic audios from this subset are capable of deceiving three widely used state-of-the-art (SOTA) SV models: ECAPA-TDNN, X-Vector, and ResNet TDNN [38, 152, 174][6].

**Methodology:** Speaker verification test sets were constructed as shown in Table 2.4. Each set consists of 20,000 randomly generated trial pairs, evenly split between positives and negatives. A positive trial pair includes two bonafide (IndicSuperb) recordings of the same target speaker, $X$, while a negative trial pair includes a bonafide (IndicSuperb) recording of $X$ with a synthetic (IndicSynth) recording of the same speaker ($X$). Since both male and female speakers contribute equally, these are referred to as combined test sets. Each combined set contains 5,000 bonafide female pairs, 5,000 bonafide male pairs, 5,000 synthetic female pairs, and 5,000 synthetic male pairs. To further analyze gender-specific performance of speaker verification systems under impersonation attacks, separate male-only and female-only test sets were also derived from the combined set.

**Evaluation Metric:** The mimicry subset is evaluated using EER. A higher EER suggests that the speaker verification model faced difficulty in distinguishing positive from negative trial pairs. This, in turn, indicates that the synthetic speech recordings in negative pairs closely resemble the target speaker's bonafide voice samples.

**Observations:** State-of-the-art speaker verification models typically report EERs below 10% on unseen language test sets [5, 181, 99]. In contrast, as shown in Table 2.4, the combined test sets from IndicSynth yield much higher EERs, ranging from 21.470% to 43.639%. Such elevated error rates suggest that the mimicry subset of IndicSynth produces speech that closely resembles the bonafide (IndicSuperb) target voices.

A comparison of male and female speaker test sets further highlights language-specific trends. For Bengali, Malayalam, and Telugu, EERs are similar across genders. In Kannada, however, female test sets consistently exhibit higher EERs than male sets, with absolute differences of 2.68% to 6.28% across verification models. This suggests that Kannada female voices in IndicSynth more closely resemble their target speakers

[6]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`

| 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 2.4: Equal Error Rates (EER%) of state-of-the-art speaker verification models under impersonation attacks using IndicSynth. Negative trial pairs consist of synthetic speech from the mimicry subset and bonafide speech of the target speakers from IndicSuperb. The elevated EER values highlight that the mimicry subset effectively resembles target voices, underscoring its utility for strengthening SV model robustness.

than Kannada male voices. Conversely, for Gujarati, Marathi, Odia, Punjabi, and Tamil, male test sets show higher EERs than female ones, indicating that male voices in these languages achieve closer mimicry of target speakers.

**Qualitative Evaluation:** To complement the quantitative results, we examined the similarity between bonafide (IndicSuperb) samples and IndicSynth's mimicry subset using t-SNE visualizations. Figures 2.6 and 2.7 show the t-SNE plots for Odia female and male speakers. For each case, we randomly selected 500 bonafide and 500 synthetic clips corresponding to the same target speakers. We then generated t-SNE plots with a perplexity value of 40, using 80-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features [108]. MFCCs are speech features inspired by the human auditory

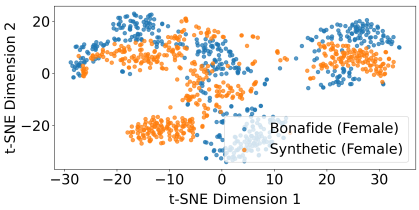



Figure 2.6: The t-SNE visualization of bonafide (IndicSuperb) Odia and IndicSynth's mimicry subset for female speakers illustrates the close proximity between bonafide and synthetic audio samples.

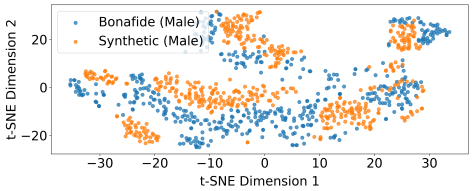



Figure 2.7: The t-SNE visualization of bonafide (IndicSuperb) Odia and IndicSynth's mimicry subset for male speakers illustrates the close proximity between bonafide and synthetic audio samples.

system. In these plots, the close proximity of bonafide and synthetic embeddings shows that the mimicry subset effectively replicates the target voices.

## 2.5 Discussion

This section outlines our motivation for creating the mimicry and diversity subsets in IndicSynth and highlights their potential applications based on experimental results.

### 2.5.1 Mimicry Subset Rationale

The mimicry subset contains synthetic voices that closely resemble the bonafide voices of target speakers. Such synthetic audios, often termed audio spoofs, can deceive speaker verification systems, enabling impersonation attacks. As demonstrated in Section 2.4.3, state-of-the-art speaker verification systems are vulnerable to these multilingual audio spoofs, underscoring the value of the mimicry subset for developing multilingual anti-spoofing solutions.

### 2.5.2 Need for a Diversity Subset

Synthetic speech is also widely misused for spreading misinformation, where the generated voices usually do not imitate specific individuals. Instead, misinformation campaigns often rely on a variety of synthetic voices. Training audio deepfake detection (ADD) models on such diverse voices is essential to improve their robustness. However, the mimicry subset is limited in diversity since it only covers voices derived from IndicSuperb speakers. To address this, we introduced the diversity subset, which extends IndicSynth with a broader range of synthetic voices beyond speaker mimicry. Table 2.1 summarizes both subsets.

### 2.5.3 Utility of the Diversity Subset

As discussed in Section 2.4.1, we benchmarked three state-of-the-art ADD models on IndicSynth (including both mimicry and diversity subsets) without domain adaptation. While these models achieve low Equal Error Rates (EERs) on ASVSpoof challenge datasets, their performance dropped significantly on IndicSynth, as evidenced by higher EERs and lower AUC scores (Table 2.2, Figure 2.4). Similar findings were reported by Munir *et al.* [108] for Urdu. These results highlight the poor generalizability of existing ADD systems to out-of-domain languages and emphasize the need for multilingual training. Combining bonafide IndicSuperb data with synthetic IndicSynth data (both mimicry and diversity subsets) can therefore provide a valuable resource for building robust, multilingual audio deepfake detection systems.

## 2.6 Conclusions and Future Work

This chapter presents IndicSynth, a large-scale multilingual synthetic speech dataset designed to advance research in multilingual audio deepfake detection and anti-spoofing. The dataset comprises approximately 4,000 hours of synthetic audio from 989 target speakers (456 female and 533 male) across 12 low-resource Indian languages. It also provides comprehensive metadata, including speaker identifiers and gender information, enabling studies on gender-related biases in audio deepfake detection and anti-spoofing.

IndicSynth is organized into mimicry and diversity subsets. The mimicry subset contains synthetic recordings (audio spoofs) that closely mimic bonafide target voices, while the diversity subset offers a broad range of realistic synthetic voices (audio deepfakes). Experimental results demonstrate that mimicry audios can successfully deceive state-of-the-art (SOTA) speaker verification systems, highlighting vulnerabilities to impersonation attacks. Similarly, SOTA audio deepfake detection models show limited performance on Indian language test sets. Both quantitative and qualitative evaluations using a SOTA language identification model confirm the linguistic authenticity of the dataset. These findings establish IndicSynth as a valuable resource for mitigating impersonation attacks and supporting robust multilingual audio deepfake detection.

This work opens several avenues for future research on linguistic biases in audio deepfake detection and anti-spoofing. For example, IndicSynth can be utilized to investigate the relatively underexplored issue of gender bias in multilingual audio deepfake detection, as well as to develop defenses against impersonation attacks involving multilingual spoofing. Furthermore, comprehensive human evaluations involving proficient speakers of the IndicSynth languages can be conducted to assess the dataset's suitability for automatic speech recognition (ASR). If found appropriate, the dataset may subsequently be explored for training ASR models. The dataset is released under the CC BY-NC 4.0 license.

## 2.7 Limitations

This chapter presents IndicSynth, a large-scale multilingual synthetic speech dataset designed to support research in multilingual audio deepfake detection and anti-spoofing. We acknowledge the following limitations:

**Scope of IndicSynth and Experimental Coverage:** IndicSynth currently includes synthetic speech recordings for 12 languages, and the mimicry subsets for Hindi and Sanskrit are not available. Future extensions could incorporate additional low-resource languages and more voice cloning models to broaden the dataset. Furthermore, our evaluations were conducted on sample test sets. While these results should reflect the overall quality of IndicSynth, more extensive evaluations could provide deeper insights.

**Lack of User Study:** Ideally, the naturalness of synthetic speech should be assessed

via a user study, with participants fluent in the target languages. However, for large-scale multilingual datasets like IndicSynth, recruiting participants proficient in low-resource languages is challenging. To meet the objectives of this study, we instead conducted experimental evaluations using state-of-the-art speaker verification models, audio deepfake detection models, and a language identification model.

The difficulty in recruiting participants familiar with low-resource languages is not unique to IndicSynth; it is a common challenge in the creation of multilingual datasets for social good. Nevertheless, with over 7,000 languages globally and increasing cases of deepfake-related misuse, there is an urgent need for multilingual synthetic datasets. The absence of user studies should not delay the release of such resources. Researchers with computational access can contribute by developing additional multilingual datasets, while collaborators with access to native speakers can later conduct user studies to evaluate human perception of synthetic speech.

Despite these limitations, the scarcity of multilingual synthetic speech datasets makes IndicSynth a valuable resource for advancing research in multilingual audio deepfake detection and anti-spoofing.

## 2.8 Ethical Considerations

Synthetic speech datasets play a crucial role in advancing research on audio deepfake detection and anti-spoofing. At the same time, we acknowledge that such datasets could potentially be misused to improve audio deepfake generation by malicious users. Hence, careful and responsible handling of these resources is essential. To this end, IndicSynth is released under the CC BY-NC 4.0 license, which prohibits commercial use. IndicSynth is built from the publicly available IndicSuperb dataset, which is licensed under CC0 (“no rights reserved”), allowing unrestricted reuse, modification, and enhancement. We strongly encourage the research community to utilize IndicSynth for socially beneficial purposes and to further the study of multilingual audio deepfake detection and anti-spoofing. Beyond its immediate utility, IndicSynth opens new avenues for research into linguistic and gender biases in audio deepfake detection and anti-spoofing. It can also support the development of defenses against multilingual spoofing attacks. The dataset is released under the CC BY-NC 4.0 license.

# CHAPTER 3

# Human Perception of IndicSynth Speech

High-quality synthetic speech technologies, including text-to-speech and voice conversion systems, generate audio that closely resembles human voices, which can be exploited for fraud, impersonation, and misinformation. This creates a pressing need for robust multilingual audio deepfake detection (ADD) and anti-spoofing systems. Chapter 2 introduced IndicSynth, a multilingual synthetic speech dataset covering 12 low-resource Indian languages. This chapter investigates human perception of IndicSynth audios to assess their naturalness, examine how convincingly the mimicry subset can imitate voices of target speakers, and determine whether the dataset presents a valuable resource for multilingual ADD and anti-spoofing research. We conduct a large-scale user study with 93 participants proficient in different Indic languages to evaluate the perceived naturalness of synthetic speech, human ability to identify bonafide recordings among collections of bonafide and synthetic audios, and the perceptual similarity of mimicry audios to target voices. The results indicate that IndicSynth audios appear highly natural (human-like), challenging to distinguish from real speech, and often capture recognizable speaker characteristics, highlighting IndicSynth's relevance for multilingual audio deepfake detection and anti-spoofing research.

## 3.1 Introduction

High-quality synthetic speech technologies, including text-to-speech (TTS) and voice conversion (VC) systems, increasingly generate audio that closely mimics human voices [8]. While these advances enable applications in assistive communication, media, and entertainment, they also amplify risks: audio deepfakes can mislead human listeners, spread misinformation, or trigger public panic, while audio spoofs can impersonate individuals and bypass speaker verification systems, enabling identity theft and fraud, as illustrated in Figure 3.1 [78, 112]. These threats are particularly concerning in linguistically diverse countries like India, where numerous low-resource languages coexist and

exposure to deepfakes is rising [64, 160]. Robust multilingual audio deepfake detection (ADD) and anti-spoofing systems are essential to counter these risks, but existing models often exhibit linguistic bias, performing poorly on out-of-domain languages [146, 145]. Developing such models requires large-scale, linguistically diverse synthetic speech datasets.


Figure 3.1: Audio deepfakes are synthetic speech recordings that sound convincingly real and can be exploited to mislead people or spread fake news, potentially causing public panic. Similarly, audio spoofs mimic a specific speaker's voice and can be misused to bypass speaker verification systems, causing impersonation and identity theft.

IndicSynth, introduced in Chapter 2, addresses this need by providing a multilingual synthetic speech dataset spanning 12 low-resource Indian languages. Before using IndicSynth to develop or evaluate ADD and anti-spoofing systems, it is crucial to understand how humans perceive the naturalness and authenticity of its synthetic audios. Human perception studies provide insights into the realism of generated speech, reveal potential vulnerabilities in speaker impersonation, and complement automated evaluations of ADD and speaker verification (SV) systems. This chapter conducts a comprehensive user study to examine whether humans can distinguish synthetic from bonafide speech, assess the naturalness of synthetic audios, and evaluate the perceptual similarity of mimicry recordings to their target speakers.

**Summary of Chapter Contributions:**

1. Conduct a large-scale user study involving 93 participants to evaluate human perception of IndicSynth audios across multiple low-resource Indian languages.
2. Measure participants' ability to distinguish bonafide versus synthetic speech and their perceived naturalness of synthetic audios using mean opinion scores (MOS).
3. Assess the perceived similarity between synthetic audios in the mimicry subset and their corresponding bonafide target voices, providing insights into speaker-specific characteristics.

## 3.2 Participant Recruitment

To conduct the user study, we recruited participants through a Google Form that outlined the purpose of the study, duration, tasks involved, incentive, and data usage policy. Participants were informed that the study aimed to evaluate the naturalness and authenticity of synthetic speech recordings. Each session lasted approximately 30 minutes, during which participants listened to audio clips and completed classification and rating tasks. Participation was voluntary, and all responses were anonymized and used solely for research purposes. An honorarium in the form of Amazon gift cards was provided as compensation. In line with ethical requirements, informed consent was obtained electronically from each participant. The study protocol was reviewed and approved by the Institutional Review Board (IRB).

The Google Form also collected basic participant information, including language proficiency and preferred time slots for participation. This enabled us to schedule sessions according to participant availability. Recruitment was carried out through email invitations, inviting participants to take part in one or two sessions. Participants who completed a single session received an honorarium of INR 150, while those who completed two sessions received INR 250. A short break of approximately 10 minutes was provided between the two sessions.

| **Language** | **# Participants** |
|---|---|
| Bengali | 5 |
| Gujarati | 4 |
| Hindi | 70 |
| Kannada | 1 |
| Malayalam | 8 |
| Marathi | 3 |
| Odia | 6 |
| Punjabi | 19 |
| Sanskrit | 2 |
| Tamil | 10 |
| Telugu | 2 |
| Urdu | 12 |
| **Total** | **93** |

Table 3.1: Participant demographics by language proficiency.

In total, 93 participants took part in the study. Some of them are proficient in multiple languages. Table 3.1 summarizes the distribution of participants across languages. Although Hindi (70) and Punjabi (19) speakers formed the majority, the study

also included participants proficient in other Indic languages such as Bengali, Gujarati, Malayalam, Marathi, Odia, Sanskrit, Tamil, Telugu, and Urdu, thereby ensuring a linguistically diverse evaluation.

## 3.3 Experimental Setup

The study involved three tasks, summarized below:

### 3.3.1 Task 1: Evaluating the Naturalness of Synthetic Speech

The first task aimed to assess the perceived naturalness of IndicSynth audio recordings. For each target language, we selected a total of 12 speech clips, comprising: three synthetic female audios, three synthetic male audios, three bonafide male audios, and three bonafide female audios. Participants were provided the test sets of only those languages in which they were proficient. Each clip had a duration of approximately 5–10 seconds. Bonafide samples were randomly selected from the IndicSuperb dataset, while synthetic samples were drawn from IndicSynth.

Participants were instructed to listen to each clip and respond to the following questions:

1. Determine whether the speech recording is real or synthetic.
2. Rate the naturalness of the audio on a 5-point Likert scale:
    - 1 – Completely unnatural (robotic or artificial)
    - 2 – Somewhat unnatural (noticeably synthetic)
    - 3 – Neutral (neither clearly natural nor unnatural)
    - 4 – Mostly natural (minor unnatural elements)
    - 5 – Completely natural (like genuine human speech)

Participants were allowed to replay each audio clip multiple times to form their judgments. This task provided quantitative measures of human-perceived naturalness, which were later analyzed using mean opinion scores (MOS) and classification accuracy. Participant responses were recorded through Google forms.

### 3.3.2 Task 2: Identifying Bonafide Audio

The second task aimed to evaluate participants' ability to identify bonafide speech among synthetic recordings. Unlike Task 1, which focused only on languages in which participants were proficient, Task 2 covered all twelve IndicSynth languages. For each language, participants were presented with one question comprising three audio clips: one generated using FreeVC24, one using XTTS-v2, and one bonafide clip from IndicSuperb. Additionally, for the Bengali question, a fourth clip generated using VITS was also included.

Participants were instructed to listen carefully to each set of audio clips and identify the bonafide clip. Responses were collected via Google Forms. The average accuracy across participants was subsequently computed and reported in the experimental results. A total of 93 participants participated in Task 2.

### 3.3.3 Task 3: Similarity Rating of Mimicry Audios

This task aimed to evaluate how closely the synthetic audios in the IndicSynth mimicry subset resemble the bonafide target voices from IndicSuperb. Due to the intensive nature of the task, which required participants to carefully listen and compare paired audio clips, this experiment was conducted on a representative subset of four IndicSynth languages: Bengali, Gujarati, Kannada, and Odia. These languages were randomly selected to cover diverse linguistic families, providing a meaningful evaluation of mimicry quality without overburdening participants.

For each language in the subset, two examples were selected: one corresponding to a male speaker and one to a female speaker. Each example included two audio clips: the original bonafide recording of the target speaker and the corresponding synthetic recording from the IndicSynth mimicry subset.

Participants were instructed to listen to both audio clips for each example and rate the similarity of the synthetic voice to the original speaker's voice on a scale of 1 to 5:

- 1: Completely dissimilar (no resemblance to the target speaker)
- 2: Somewhat dissimilar (few similarities but mostly different)
- 3: Neutral (equal mix of similarities and differences)

- 4: Mostly similar (closely resembles the target speaker)
- 5: Very similar (indistinguishable from the target speaker)

A total of 92 participants completed this task. The similarity ratings collected from this representative subset provide reliable insights into the human-perceived mimicry quality of IndicSynth audios.

## 3.4 Experiments and Results

### 3.4.1 Task 1: Real vs. Synthetic Classification

The goal of this experiment is to examine whether human listeners can reliably distinguish between bonafide and synthetic speech. As explained in Section 3.3.1, participants listened to a balanced set of real and generated clips in their proficient languages and completed both a binary classification and a naturalness rating task.

We presented each participant with 12 clips per language, spanning male and female speakers from both real and synthetic sources. We collected two types of responses: (i) binary classification for authenticity, and (ii) a naturalness score on a 5-point Likert scale. Accuracy served as a direct indicator of human ability to detect synthetic speech, while MOS captured the perceptual realism of both real and synthetic audios. These two metrics complement each other: accuracy measures detection ability, whereas MOS measures naturalness score.

**Observations.** As illustrated in the Table 3.2, the average classification accuracy across all languages and models was **53.96%**, which indicates random guessing in classification. This result shows that the participants struggled to accurately detect the synthetic audios. Accuracy varied across languages: Hindi (67.5%) and Odia (62.5%) were comparatively higher, while Marathi (36.1%) and Kannada (41.7%) were much lower. XTTS-v2 generally achieved slightly higher accuracies than FreeVC24, except in a few languages such as Marathi and Punjabi where FreeVC24 outperformed XTTS-v2.

MOS scores illustrated in the Table 3.3 reveal a similar trend. The average MOS for real clips and synthetic clips are mostly very close, with differences typically below one point. This indicates that participants found synthetic clips nearly as natural as

| Language | Model | Accuracy (%) |
|---|---|---|
| Bengali | XTTS-v2 | 58.333 |
| Bengali | VITS | 61.667 |
| **Bengali** | **FreeVC24** | **40.000** |
| Gujarati | XTTS-v2 | 56.250 |
| **Gujarati** | **FreeVC24** | **43.750** |
| Hindi | XTTS-v2 | 67.500 |
| **Hindi** | **FreeVC24** | **58.214** |
| **Kannada** | **XTTS-v2** | **41.667** |
| Kannada | FreeVC24 | 50.000 |
| **Malayalam** | **XTTS-v2** | **44.792** |
| Malayalam | FreeVC24 | 53.125 |
| **Marathi** | **XTTS-v2** | **36.111** |
| Marathi | FreeVC24 | 58.333 |
| Odia | XTTS-v2 | 62.500 |
| **Odia** | **FreeVC24** | **61.111** |
| **Punjabi** | **XTTS-v2** | **53.947** |
| Punjabi | FreeVC24 | 61.403 |
| Sanskrit | XTTS-v2 | 62.500 |
| **Sanskrit** | **FreeVC24** | **54.167** |
| Tamil | XTTS-v2 | 55.833 |
| **Tamil** | **FreeVC24** | **52.500** |
| Telugu | XTTS-v2 | 58.333 |
| **Telugu** | **FreeVC24** | **54.167** |
| Urdu | XTTS-v2 | 56.944 |
| **Urdu** | **FreeVC24** | **45.833** |
| **Average** | **-** | **53.959** |

Table 3.2: Average participant accuracy (%) in the audio deepfake detection task across languages and synthesis models. For each language, the row with the lowest average accuracy is highlighted.

real ones. Among the models, FreeVC24 consistently produced synthetic audios whose MOS scores were closest to the bonafide clips, highlighting that FreeVC24 generates highly realistic synthetic audios. In some languages such as Bengali, Gujarati, and Urdu for FreeVC24, the MOS of synthetic clips even exceeded that of the real clips. A similar reversal occurred for Kannada, Malayalam, and Marathi in XTTS-v2. These cases suggest that IndicSynth's synthetic clips occasionally sounded more realistic than original recordings.

**Key Insights:** We summarize our key insights from this experiment below:

1. Human listeners found it difficult to differentiate between synthetic and bonafide clips, as reflected in near-random classification accuracy.
2. The small MOS gap between real and synthetic clips shows that IndicSynth audios sound convincingly natural across languages.
3. FreeVC24 generated more realistic audios than XTTS-v2 for most languages, often producing MOS scores indistinguishable from real speech.

| Language | Model | Avg. MOS: Real Clips | Avg. MOS: Synthetic Clips | MOS Difference (Real − Fake) |
|---|---|---|---|---|
| Bengali | XTTS-v2 | 3.467 | 3.233 | **0.234** |
| Bengali | VITS | 3.067 | 2.267 | 0.800 |
| **Bengali** | **FreeVC24** | **2.900** | **3.500** | **-0.600** |
| Gujarati | XTTS-v2 | 3.292 | 3.250 | 0.042 |
| **Gujarati** | **FreeVC24** | **2.417** | **2.792** | **-0.375** |
| Hindi | XTTS-v2 | 3.762 | 2.876 | 0.496 |
| Hindi | FreeVC24 | 3.511 | 3.102 | **0.409** |
| **Kannada** | **XTTS-v2** | **2.667** | **3.500** | **-0.833** |
| Kannada | FreeVC24 | 3.667 | 3.333 | 0.334 |
| **Malayalam** | **XTTS-v2** | **2.750** | **2.917** | **-0.167** |
| Malayalam | FreeVC24 | 2.938 | 2.854 | 0.084 |
| **Marathi** | **XTTS-v2** | **3.500** | **4.222** | **-0.722** |
| Marathi | FreeVC24 | 3.833 | 3.333 | 0.500 |
| Odia | XTTS-v2 | 4.028 | 3.250 | 0.778 |
| Odia | FreeVC24 | 4.056 | 3.611 | **0.445** |
| Punjabi | XTTS-v2 | 3.421 | 3.184 | **0.237** |
| Punjabi | FreeVC24 | 3.421 | 2.859 | 0.562 |
| Sanskrit | XTTS-v2 | 3.583 | 3.333 | 0.250 |
| Sanskrit | FreeVC24 | 3.333 | 3.083 | 0.250 |
| Tamil | XTTS-v2 | 3.317 | 2.933 | 0.384 |
| Tamil | FreeVC24 | 3.383 | 3.367 | **0.016** |
| Telugu | XTTS-v2 | 3.167 | 2.333 | 0.834 |
| Telugu | FreeVC24 | 3.083 | 2.667 | **0.416** |
| Urdu | XTTS-v2 | 3.403 | 3.056 | 0.347 |
| **Urdu** | **FreeVC24** | **2.917** | **3.250** | **-0.333** |
| **Average** | - | **3.315** | **3.124** | **0.191** |

Table 3.3: Average naturalness score indicated through MOS on a Likert scale. Higher score indicates higher naturalness of audio clips as perceived by humans. For each language, cell with the least MOS difference is highlighted in Blue. Rows with a negative MOS difference are highlighted in Red.

4. In some cases, synthetic clips outscored real ones on MOS. The presence of occasional disfluencies in generated speech contributed to its human-like quality, making it appear natural to listeners despite being synthetic.

5. The results emphasize both the strength and risk of modern generative models: while they enable realistic multilingual synthesis, they also increase the challenge of reliable human and machine detection, leading to a threat of audio deepfake-related frauds.

### 3.4.2 Experiment 2: Identifying Bonafide Audio

The goal of this experiment is to evaluate whether participants can reliably identify bonafide recordings when presented alongside synthetic audios generated by different speech synthesis models. This task directly reflects the perceptual difficulty of distin-

guishing authentic speech from deepfakes.

As described in Section 3.3.2, each participant was presented with 12 multiple-choice questions, one for each IndicSynth language. Each question contained one bonafide clip from IndicSuperb and corresponding synthetic clips generated by FreeVC24 and XTTS-v2 (with an additional VITS clip for Bengali). Participants were instructed to select the bonafide clip in each case. A total of 93 participants completed this task.

**Observations.** We computed the accuracy of each participant over 12 questions and then averaged across all participants. The overall average accuracy was found to be **42.83%**, which is close to random guess, highlighting the high perceptual realism of the IndicSynth speech samples and the challenges posed in bonafide identification.

### 3.4.3 Task 3: Similarity Rating of Mimicry Audios

The aim of this experiment is to evaluate the perceived similarity between synthetic audios in the IndicSynth mimicry subset and the corresponding bonafide target voices. This task helps quantify how closely synthetic speech captures speaker-specific characteristics from real recordings.

As described in Section 3.3.3, participants listened to paired audio clips—one bonafide and one synthetic—from a representative subset of four IndicSynth languages: Bengali, Gujarati, Kannada, and Odia. Each participant rated the similarity of the synthetic voice to the original speaker on a 1–5 Likert scale. A total of 92 participants completed this task, and the ratings were averaged per language to compute the mean opinion score (MOS).

**Observations.** Table 3.4 reports the average MOS for each language. All languages scored above 3, with Bengali and Gujarati at 3.35 and 3.33, Kannada slightly higher at 3.51, and Odia at 3.07. These consistently above-neutral scores indicate that participants clearly noticed similarities between the synthetic audios and the original speakers' voices. This demonstrates that IndicSynth's mimicry subset effectively captures speaker-specific characteristics, producing synthetic speech that is perceptually close to the target voices.

**Key Insights:**

| Language | Average MOS |
|---|---|
| Bengali | 3.348 |
| Gujarati | 3.326 |
| Kannada | 3.511 |
| Odia | 3.065 |
| **Average** | **3.312** |

Table 3.4: Mean opinion scores (MOS) indicating perceived similarity between synthetic and original voices for the mimicry subset.

1. The MOS scores above 3 indicate that participants perceived a moderate similarity between synthetic voices and the bonafide voice recordings of original speakers, suggesting that the mimicry subset captures recognizable speaker characteristics.

2. Higher similarity scores for Kannada suggest relatively closer speaker modeling, while slightly lower scores for Odia reflect perceptible differences, though still above neutral.

3. These findings highlight the utility of IndicSynth's mimicry subset for anti-spoofing research, as it provides synthetic voices that are decently similar to real voices to test the robustness of speaker verification and audio deepfake detection systems.

## 3.5 Conclusions and Future Work

This chapter describes the comprehensive user study that was conducted to evaluate human perception of synthetic speech present in IndicSynth. The study was conducted with a total of 93 human participants. Participants performed tasks involving naturalness ratings, real versus synthetic classification, bonafide audio identification, and similarity assessment of mimicry audios across multiple Indic languages. The study revealed that human listeners generally struggled to reliably distinguish synthetic speech from bonafide recordings, with classification accuracies close to random guessing. Mean opinion scores further indicated that synthetic audios were perceived as nearly as natural as real speech, and in some cases even more so, particularly when generation-related disfluencies contributed to human-like qualities. The similarity ratings in the mimicry subset demonstrated that synthetic voices effectively capture speaker-specific characteristics.

These findings highlight both the potential and the risks associated with high-quality synthetic speech. The high naturalness of synthetic audios increases the likelihood of audio deepfake–related fraud, while the strong similarity between bonafide and syn-

thetic audios amplifies the threat of spoofing attacks. Addressing these risks requires robust multilingual audio deepfake detection and anti-spoofing systems. Large-scale, linguistically diverse synthetic speech datasets are essential for training such systems, and IndicSynth provides a valuable resource for multilingual anti-spoofing and audio deepfake detection research.

Future work can expand the scope of the study to include a larger and more diverse participant pool, additional languages, and a wider range of synthesis models. Incorporating more extensive evaluation metrics that could provide deeper insights into human perception of synthetic speech. Longitudinal studies investigating listener adaptation to synthetic voices, as well as strategies for improving detection robustness in real-world scenarios, would further enhance the practical utility of the dataset.

## 3.6 Limitations

Despite the insights gained from the user study, several limitations constrain the generalizability of the findings. First, the participant pool, while linguistically diverse, was heavily skewed toward Hindi and Punjabi speakers, resulting in limited representation for other Indic languages. Consequently, the results may not fully capture perceptions across all low-resource languages included in IndicSynth. Second, participants' fluency in the tested languages was self-reported, which may have introduced variability in comprehension and perceptual judgments. Third, the scale of the study was constrained by practical considerations, with each participant evaluating only a subset of the total audio clips.

# CHAPTER 4

# Cross-Task Profiling of Speech Datasets

The growing demand for inclusive speech technologies has amplified the need for multilingual datasets for research. To address this need, Chapter 2 introduced IndicSynth, and Chapter 3 illustrated human perception of the dataset. While IndicSynth is designed specifically for multilingual audio deepfake detection and anti-spoofing tasks, several other speech-based tasks continue to remain underserved due to resource limitations. Furthermore, limited awareness of existing task-specific datasets continues to hinder research progress, particularly in linguistically diverse countries such as India. One promising strategy to address this challenge is cross-task profiling, which involves evaluating the usability of existing datasets across multiple downstream tasks, rather than restricting them to their originally intended scope. While most surveys catalog available datasets, the cross-task utility of these datasets remains underexplored.

To fill this gap, this chapter introduces *Task-Lens*[1], a cross-task profiling framework that assigns a task-specific utility score to quantify the relevance of a dataset for various speech processing tasks. Task-Lens is applied to evaluate the utility of 34 publicly available Indian speech datasets covering 26 languages across eight downstream tasks: Automatic Speech Recognition (ASR), Audio Deepfake Detection (ADD), Emotion Recognition, Gender Recognition, Language Identification (LID), Multilingual Text-to-Speech (TTS), Monolingual TTS, and Speaker Verification/Identification. This evaluation includes IndicSynth, allowing us to further investigate its potential utility across tasks beyond its original design for ADD and anti-spoofing research.

This chapter presents our findings on the usability of these datasets across tasks, identifies areas for improvement to enhance their utility, and highlights downstream tasks that remain critically under-resourced. The insights aim to assist researchers in finding relevant datasets for their research problems and guiding future data development efforts, ultimately supporting the creation of multilingual speech datasets that are broadly applicable across diverse tasks.

[1]Part of this chapter has been accepted for publication at LREC 2026:
Swati Sharma, Divya V. Sharma, Anubha Gupta. 2025. Task-Lens: Cross-Task Utility Based Speech Dataset Profiling for Low-Resource Indian Languages.

## 4.1 Introduction

The growing demand for inclusive speech technologies has intensified the need for multilingual datasets to support research in Natural Language Processing (NLP). However, most publicly available speech datasets are predominantly English-centric, which restricts progress in developing speech technologies for low-resource languages. While multilingual speech models aim to improve inclusivity, they frequently underperform and exhibit linguistic biases, particularly when evaluated on underrepresented languages [186]. This underscores the critical need for task-specific datasets in low-resource settings.

In addition to creating new datasets, optimizing the use of existing resources presents an effective strategy for alleviating data scarcity. However, researchers often remain unaware of publicly available datasets for underrepresented languages, further impeding research progress for those languages [85]. In this context, cross-task profiling emerges as a promising approach. Cross-task profiling involves investigating a dataset's utility across multiple downstream tasks, rather than solely using it for its originally intended purpose. Such an approach can surface the untapped potential of existing resources and shed light on both research and dataset development priorities.

Despite its potential, cross-task profiling remains underexplored. Prior work in the field has largely focused either on dataset creation methodologies or on cataloging existing datasets [149, 11]. Even when datasets are accompanied by rich metadata, they are typically positioned as relevant to only a single task [156, 67, 66, 50]. As a result, the broader task-specific applicability of speech datasets remains unclear, especially for low-resource languages. While this challenge is global, it is particularly pressing in linguistically diverse countries such as India, where maximizing the utility of existing datasets can significantly advance multilingual speech research.

To address this gap, this chapter introduces Task-Lens, a cross-task utility-based profiling framework for speech datasets, illustrated in Figure 4.1. Task-Lens computes a task-specific utility score for each dataset, thereby facilitating systematic cross-task profiling. We apply Task-Lens to evaluate 34 publicly available Indian speech datasets spanning 26 languages and covering over 74,745 hours of audio across eight downstream tasks: Automatic Speech Recognition (ASR), Audio Deepfake Detection

(ADD), Emotion Recognition, Gender Recognition, Language Identification (LID), Multilingual Text-to-Speech (TTS), Monolingual TTS, and Speaker Verification/Identification. Through Task-Lens based cross-task profiling of the target Indian speech datasets across the eight downstream tasks, we address the following key research questions:

- Which datasets are currently usable for which tasks?
- What improvements would enhance a dataset's suitability for additional tasks?
- Which speech tasks currently lack dataset support?
- Which Indian languages are adequately represented for each speech task?
- Which language-task combinations currently exhibit significant resource gaps?

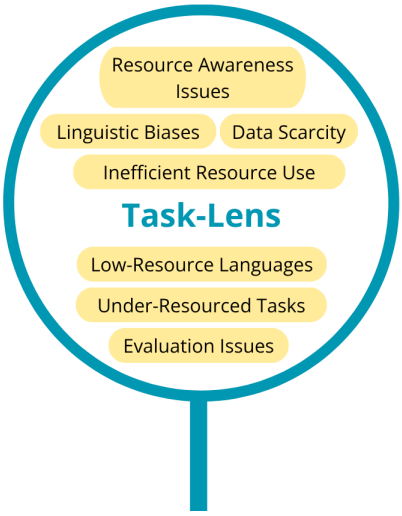



Figure 4.1: Addressing Key Barriers to Inclusive Speech Technology Development through Task-Lens

**Summary of Chapter Contributions:**

1. We propose **Task-Lens**, a cross-task utility profiling framework that assigns task-specific utility scores to quantify the utility of speech datasets for diverse downstream tasks.

2. We apply Task-Lens on **34 publicly available Indian speech datasets**, spanning **26 languages** and over **74,745 hours** of audio for a comprehensive cross-task profiling across **eight speech processing tasks**: *Automatic Speech Recognition (ASR), Audio Deepfake Detection (ADD), Emotion Recognition, Gender Recognition, Language Identification (LID), Multilingual Text-to-Speech (TTS), Monolingual TTS, and Speaker Verification/Identification*.

3. Based on the analysis, we present empirical insights into:
   - Dataset-task compatibility and coverage
   - Task-specific limitations of existing datasets

- Task-language combinations that are under-resourced
- Opportunities to enhance the cross-task utility of existing speech datasets
- Suggest directions for future data collection efforts

## 4.2 Related Works

### 4.2.1 Cross-Task Utility Evaluation

Several studies have explored inclusive profiling frameworks aimed at evaluating language representations across multiple tasks and languages [191, 34, 49, 22, 100]. However, the majority of surveys and benchmarks primarily focus on compiling lists of available datasets or assessing the utility of newly introduced resources [149, 11]. As a result, a number of important research questions remain insufficiently addressed. For example: (1) Although datasets are often released for a particular task, their rich metadata may enable broader applicability — how can we systematically examine such cross-task potential for existing multilingual speech datasets? (2) Are there urgent gaps for certain language–task combinations that require immediate dataset creation? Addressing these questions could reveal critical gaps in NLP research. Overall, cross-task evaluation from an inclusivity perspective remains largely underexplored, underscoring the need for unified, utility-driven frameworks, particularly for linguistically diverse countries such as India.

### 4.2.2 Indian Dataset Landscape

India, with its 22 constitutionally recognized languages, remains underrepresented in mainstream speech benchmarks [69]. High-quality, general-purpose datasets that capture dialectal and accent variations are especially scarce [16]. While some targeted resources exist — such as Lahaja for Hindi accents, AccentDB for non-native Indian English, and Svarah for speech from 19 Indian states [67, 3, 66] — awareness of these datasets is limited. The most comprehensive review of Indian speech datasets across diverse tasks was published over a decade ago by Pukhraj P. Shrishrimal [128]. The absence of up-to-date consolidated reviews means that researchers often spend significant time locating relevant datasets instead of advancing novel research. To bridge this gap,

this chapter introduces *Task-Lens*, a framework designed to profile the cross-task utility of Indian speech datasets and identify critical resource gaps.

## 4.3 Task-Lens

Task-Lens is a utility-driven, cross-task profiling framework designed for systematic evaluation of speech datasets. The framework operates in four sequential stages: *dataset discovery*, *dataset filtering*, *feature extraction*, and *utility mapping*. An overview of the process is presented in Figure 4.2.

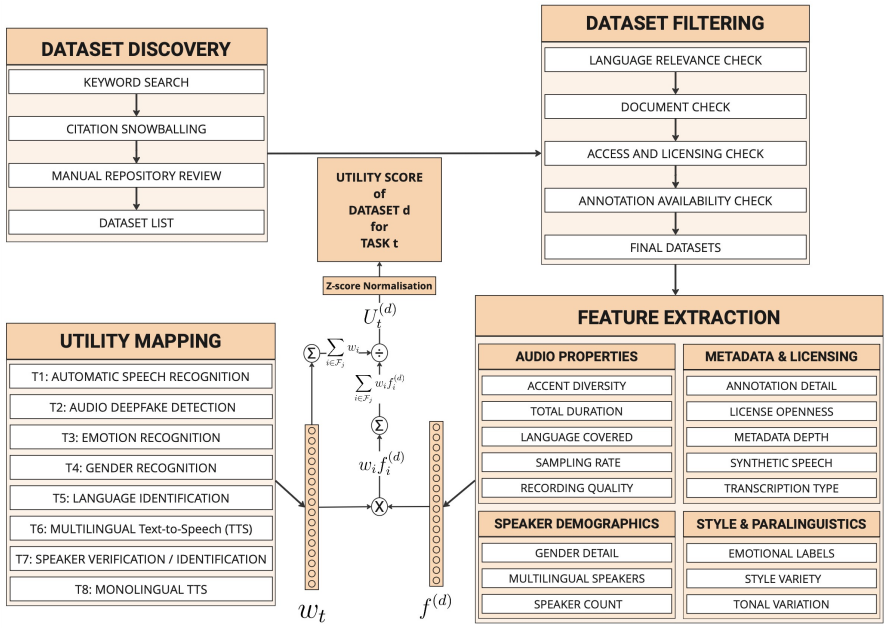



Figure 4.2: Overview of the Task-Lens framework. The process comprises four stages: dataset discovery, dataset filtering, feature extraction, and utility score computation. Through utility mapping, dataset features are aligned with the requirements of specific downstream tasks to assess suitability. The utility score $U_t^{(d)}$ measures the relative suitability of dataset $d$ for task $t$, based on task-specific weights $w_t$ and dataset features $f^{(d)}$. The framework supports eight speech processing tasks: Automatic Speech Recognition [T1], Audio Deepfake Detection [T2], Emotion Recognition [T3], Gender Recognition [T4], Language Identification [T5], Multilingual Text-to-Speech [T6], Speaker Verification/Identification [T7], and Monolingual Text-to-Speech [T8].

### 4.3.1 Dataset Discovery

The initial stage focuses on identifying relevant datasets for inclusion in the study. A systematic search was conducted on Google Scholar, arXiv, GitHub, Hugging Face,

and PapersWithCode. The search employed a set of targeted keywords, such as Indian speech dataset, Hindi ASR corpus, Indian multilingual TTS, Indian language speech data, Bhasha ASR dataset, low-resource Indian speech corpus, Indian speech translation dataset, and Indic voice dataset. To further expand coverage, citation snowballing was also applied. This process resulted in a curated list of Indian speech datasets for subsequent analysis.

### 4.3.2 Dataset Filtering

After identifying potential datasets, the next step involves filtering them based on their authenticity and accessibility. Therefore, each dataset's repository was manually reviewed to record its name, access link, and license information. Only datasets meeting all of the following criteria were retained:

1. Contain speech data from the Indian demographic population.
2. Include basic documentation—such as a research paper, preprint, metadata file, or README—describing the dataset's contents and collection methodology.
3. Provide a functional access link and a valid license.
4. Include metadata with at least one type of annotation, for example, utterance transcripts, speaker identifiers, or language tags.

Datasets were excluded if they had any of the following issues: inactive access link, absence of speech audio, lack of documentation, or duplication as an exact copy or superseded release.

Applying these criteria resulted in a curated set of 34 publicly available Indian speech datasets. Table 4.1 presents the complete list, including dataset abbreviations and references. Collectively, these datasets span 26 Indian languages, totaling over 74,745 hours of audio. They encompass both monolingual and multilingual resources, representing a wide range of speech styles, domains, speaker demographics, and levels of annotation detail.

| ID | Dataset |
|---|---|
| $D_1$ | AccentDB [3] |
| $D_2$ | Assamese TTS Corpus [163] |
| $D_3$ | BhasaAnuvaad [63] |
| $D_4$ | Open-source TTS Voices [154] |
| $D_5$ | Bengali Numbers Corpus [115] |
| $D_6$ | South Asian Crowdsourced Speech [81] |
| $D_7$ | Low-Income Workers [1] |
| $D_8$ | FLEURS [34] |
| $D_9$ | GACMIS Songs [168] |
| $D_{10}$ | GlobalPhone Speaker Package [43] |
| $D_{11}$ | GRAM VAANI [14] |
| $D_{12}$ | Hindi-Tamil-English ASR [61] |
| $D_{13}$ | Indian Folk Music [151] |
| $D_{14}$ | Regional Music [150] |
| $D_{15}$ | Indic TTS (IITM) [62] |
| $D_{16}$ | IndicSpeech (TTS) [156] |
| $D_{17}$ | IndicSUPERB [65] |
| $D_{18}$ | IndicVoices-R [141] |
| $D_{19}$ | Kashmiri Data Corpus [121] |
| $D_{20}$ | KritiSamhita [82] |
| $D_{21}$ | Lahaja [67] |
| $D_{22}$ | MS Indic Speech [104] |
| $D_{23}$ | MUCS Site [40] |
| $D_{24}$ | Nexdata AI 759h [117] |
| $D_{25}$ | NISP [75] |
| $D_{26}$ | NPTEL2020 [4] |
| $D_{27}$ | Opensource Multispeaker Data [56] |
| $D_{28}$ | Rajasthani Hindi (MS) [106] |
| $D_{29}$ | Rasa [155] |
| $D_{30}$ | SMC Malayalam [33] |
| $D_{31}$ | Svarah [66] |
| $D_{32}$ | Urdu Recognition (Desktop) [44] |
| $D_{33}$ | Vāksañcayaḥ [2] |
| $D_{34}$ | IndicSynth (Ours) |

Table 4.1: List of the 34 publicly available Indian speech datasets analyzed in this study, each assigned a unique identifier ($D_1$–$D_{34}$) for consistent reference. The list comprises both prominent and lesser-known resources, with corresponding citations provided.

### 4.3.3 Feature Extraction

Following dataset filtering, 16 descriptive features are extracted for each dataset. The selection is informed by an analysis of metadata commonly available in speech datasets. Table 4.2 summarizes these features and explains the rationale behind each. They are grouped into four dimensions:

1. **Audio properties:** Accent Diversity (number of distinct accents/dialects); Total duration (hours); Language Covered (number of unique languages represented); Sampling rate (kHz); Recording Quality (studio, mixed, or noisy)
2. **Metadata and licensing:** Annotation detail (utterance, word, phoneme); License Openness (e.g., CC0, CC-BY); Metadata depth (presence of speaker ID, demographic attributes); Synthetic speech (binary); Transcription type (normalized and aligned, verbatim, noisy).
3. **Speaker demographics:** Gender detail (presence of male and female speakers); Multilingual speakers (average number of languages per speaker); Speaker count.
4. **Style and paralinguistics:** Emotion labels (unique label count); Style variety (read, conversational, spontaneous); Tonal variation (unique label count).

| **Feature** | **Rationale** |
|---|---|
| $f_1$: Accent Diversity | Phoneme coverage |
| $f_2$: Annotation Detail | Linguistic modeling |
| $f_3$: Emotion Labels | Affective computing |
| $f_4$: Gender Detail | Fairness in gender tasks |
| $f_5$: License Openness | Reuse and distribution |
| $f_6$: Languages Covered | Multilingual generalization |
| $f_7$: Multilingual Speakers | Cross-lingual adaptability |
| $f_8$: Metadata Depth | Demographic analysis |
| $f_9$: Recording Quality | Environmental variability |
| $f_{10}$: Sampling Rate | Audio fidelity |
| $f_{11}$: Speaker Count | Speaker variability |
| $f_{12}$: Style Variety | Prosody modeling |
| $f_{13}$: Synthetic Speech | Data Augmentation |
| $f_{14}$: Transcription Type | ASR/TTS alignment |
| $f_{15}$: Total Duration | Data scale |
| $f_{16}$: Tone Variation | Intonation modeling |

Table 4.2: Utility feature summary.

Each feature score $f_i^{(d)}$ is derived using a feature-specific computation and scaled to the [0,5] range, as specified in Table 4.3. The process employs two main approaches: (1) z-score normalization for continuous features to represent their relative magnitude across datasets, and (2) threshold-based categorical scoring to represent qualitative differences through clear, discrete levels. This combined approach provides a consistent framework for evaluating both quantitative and structural aspects of datasets.

| **Feature ($f_i$)** | **Computation Description** |
|---|---|
| $f_1, f_6, f_{11}, f_{15}$ | Z-score-based normalization relative to the mean and standard deviation across datasets; result is bounded between 0 and 5. |
| $f_2$ | Assign 5 if annotation is at phoneme level, 3 for word level, 1.5 for utterance or sentence level, and 0 otherwise (finer granularity improves ASR model training and error analysis, as applied in our ASR evaluation). |
| $f_3$ | Let $t$ be the number of distinct tags. Assign 5 if $t \geq 6$, 3 if $2 \leq t \leq 5$, 1 if $t = 1$, and 0 otherwise. |
| $f_4$ | Assign 5 if both male and female speakers are present, 3 if only one gender is present, 0 if none. |
| $f_5$ | Assign 5 for CC0 license, 4 for CC-BY/MIT/Apache, 3 for GPL/CC-SA/unspecified CC, 2 for non-commercial, 1 for restricted/custom/commercial, and 0 otherwise. |
| $f_7$ | Let $n$ be the number of languages per speaker. Assign 5 if $n \geq 3$, 3 if $n \geq 2$, 1 if $n \geq 1$, and 0 otherwise. |
| $f_8$ | Let $m$ be the number of informative metadata fields (e.g., speaker age, gender, dialect, region). Assign 5 if $m > 4$, 3.5 if $1 \leq m \leq 3$, 2 if $m = 1$, 0 otherwise. |
| $f_9$ | Assign 5 for 'noise', 4 for 'mixed', 2.5 for 'clean', 0 for anything else. |
| $f_{10}$ | Let $s$ be the set of all sampling rates. Assign 5 if all $s_i \geq 48$kHz, 3.5 if all $s_i = 32$kHz and some $s_i < 48$kHz, 3 if any $s_i < 32$kHz, 1.5 if 8kHz is present, and 0 otherwise. |
| $f_{12}$ | Assign 5 if all three styles (read/conversational/spontaneous) are present, 4 if any two of those or if music is present, 2 if only spontaneous, 1 if only read, 0 otherwise. |
| $f_{13}$ | Assign 5 if synthetic speech is available, otherwise 0. |
| $f_{14}$ | Assign 5 if transcript is phonemic or normalized, 3 if verbatim/clean, 0 otherwise. |
| $f_{16}$ | Let $t$ be the number of distinct tags. Assign 5 if $t \geq 4$, 3 if $1 \leq t \leq 5$, 1 if $t = 1$, and 0 otherwise. |

Table 4.3: Feature-specific scoring $f_i$ is applied to normalize each dataset attribute onto the [0,5] scale. As summarized in the table, this involves z-score normalization for continuous features and threshold-based categorical mapping for qualitative features, ensuring consistent and interpretable representation across varied speech datasets.

### 4.3.4 Utility Mapping

Upon completing feature extraction, each dataset is described by its full set of features. The subsequent step involves mapping these features to task-specific utility scores, considering eight representative speech technology tasks:

$T_1$ : Automatic Speech Recognition (ASR),

$T_2$ : Audio DeepFake Detection (ADD),

$T_3$ : Emotion Recognition,

$T_4$ : Gender Recognition,

$T_5$ : Language Identification (LID),

$T_6$ : Multilingual text-to-speech,

$T_7$ : Speaker Verification/Identification,

$T_8$ : Monolingual Text-to-Speech (TTS),

$T_6$ and $T_8$ are considered separately to capture the distinction between generating speech across multiple languages and producing high-quality output in a single language[2].

For each task $t$, let $\mathcal{F}_t \subseteq \{f_1, \ldots, f_{16}\}$ denote the subset of utility features identified as relevant, each assigned a positive importance weight $w_i > 0$. The weights for each task are provided in Table 4.4, determined by both the functional demands of the task and the empirical significance of features in prior speech research. For example, ADD places higher weight on synthetic data availability, emotion recognition focuses on emotion labels, and ASR prioritizes recording and transcription quality. The assigned weights indicate the extent to which each feature affects a downstream task, and they can be adapted by researchers as requirements evolve.

The utility score of a dataset $d$ for a given task $t$ is calculated and subsequently standardized through z-score normalization:

$$\widetilde{U}_t^{(d)} = \frac{\sum_{i \in \mathcal{F}_\sqcup} w_i \, f_i^{(d)}}{\sum_{i \in \mathcal{F}_\sqcup} w_i}, U_t^{(d)} = \frac{\widetilde{U}_t^{(d)} - \mu_t}{\sigma_t}, \tag{4.1}$$

where $\mu_t$ and $\sigma_t$ are the mean and standard deviation of utility values across all datasets for task $t$. We denote $U_t^{(d)}$ as the Utility Score for task $t$ and dataset $d$. The standardized utility scores for all tasks are then aggregated to form a multi-task utility profile:

[2]For $T_8$, feature $f_{15}$ is modified to represent the maximum duration (in hours) of any single language, rather than the total duration of the entire dataset, thus emphasizing performance in one language.

| **Feature** ($f_i$) | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ |
|---|---|---|---|---|---|---|---|---|
| $f_1$ | 2.50 | 3.25 | 2.50 | 4.00 | 3.75 | 3.25 | 4.00 | 1.75 |
| $f_2$ | 4.00 | 0.00 | 3.75 | 1.00 | 1.75 | 3.75 | 0.00 | 2.75 |
| $f_3$ | 0.00 | 0.00 | 5.00 | 2.00 | 1.25 | 1.00 | 1.50 | 1.25 |
| $f_4$ | 2.25 | 3.75 | 2.75 | 5.00 | 4.00 | 3.75 | 3.50 | 2.50 |
| $f_5$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $f_6$ | 1.75 | 2.75 | 2.25 | 3.00 | 5.00 | 5.00 | 4.25 | 0.00 |
| $f_7$ | 0.00 | 0.00 | 0.50 | 1.50 | 1.50 | 3.40 | 3.75 | 0.00 |
| $f_8$ | 0.00 | 0.00 | 1.50 | 1.25 | 2.25 | 1.50 | 2.25 | 0.75 |
| $f_9$ | 3.50 | 2.25 | 3.25 | 3.75 | 3.50 | 4.25 | 4.75 | 3.25 |
| $f_{10}$ | 5.00 | 1.75 | 1.25 | 3.50 | 3.25 | 3.00 | 4.50 | 3.50 |
| $f_{11}$ | 2.75 | 4.25 | 4.25 | 4.50 | 4.25 | 4.50 | 5.00 | 4.00 |
| $f_{12}$ | 2.00 | 0.00 | 2.00 | 3.25 | 2.50 | 2.50 | 2.50 | 2.00 |
| $f_{13}$ | 0.00 | 5.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.25 |
| $f_{14}$ | 4.50 | 0.00 | 0.00 | 1.00 | 0.75 | 4.00 | 0.00 | 3.00 |
| $f_{15}$ | 3.00 | 1.25 | 1.75 | 3.50 | 3.00 | 3.59 | 3.25 | 5.00 |
| $f_{16}$ | 0.00 | 0.00 | 0.25 | 2.00 | 1.25 | 2.75 | 1.25 | 1.00 |

Table 4.4: Feature importance weights ($w_i$ ) for each task used in utility computation. Each feature $f_i$ is assigned a weight between 0 and 5, indicating its relevance to the specific requirements of each downstream task $T_j$. These weights serve as the basis for calculating dataset utility scores across various speech technology tasks.

$$\mathbf{U}^{(d)} = [U_1^{(d)}, \ldots, U_8^{(d)}]^\top \in \mathbb{R}^8.$$

This z-score–based approach allows for consistent dataset comparisons by representing each utility score as its deviation from the mean for that task. Higher values reflect stronger alignment with the priorities of the given task. The framework further facilitates automated dataset selection and benchmarking, integrating smoothly into evaluation pipelines to enhance reproducibility and enable task-oriented dataset curation.

## 4.4 Task-Lens Utility Evaluation

### 4.4.1 Cross-Task Dataset Utility

In speech processing, it is common to design each dataset $d$ with a single target task $t$ in mind. However, many released datasets include extensive metadata, making them applicable to a wider range of tasks. For example, while Librispeech was originally created for ASR, it has also been used for speaker verification [122, 147]. Despite this

potential, the cross-task utility of existing datasets remains largely underexplored. As a result, without systematic evaluation of their cross-task value, datasets are often underutilized for varied applications. This gap is especially challenging for NLP researchers working with low-resource languages. To address this, we employ Task-Lens to investigate: *(1) How well does each dataset $d$ serve different tasks $t$? (2) What improvements could enhance a dataset's relevance for cross-task use?*

**Setup:** Task-Lens addresses these questions by calculating a normalized utility score $U_t^{(d)}$ for each dataset $d$ with respect to every task $t$, as shown in Table 4.5. A higher $U_t^{(d)}$ reflects greater suitability of dataset $d$ for task $t$. The computation of these utility scores follows the procedure in Equation 4.1 (Section 4.3). Details about the tasks and datasets can be found in Section 4.3 and Table 4.1.

**Observations:** As presented in Table 4.5, datasets $D_8$, $D_{17}$, $D_{18}$, and $D_{34}$ exhibit the highest cross-task utility, each achieving a utility score ($U_t^{(d)}$) above 1.0 for most tasks. In particular, $D_{18}$ (IndicVoices-R) and $D_{34}$ (IndicSynth) reaches $U_t^{(d)} \geq 1.25$ for the majority of tasks.

**Insights:** To our knowledge, this review captures the majority of known feature – dataset associations; however, certain features were absent in specific datasets: $f_1$ in $d_3$, $d_7$, $d_8$, $d_9$, $d_{10}$; $f_2$ in $d_1, d_{20}, d_{32}$; $f_4$ in $d_{10}$; $f_6$ in $d_{21}$; $f_{10}$ in $d_{12}$; $f_{11}$ in $d_{12}, d_{22}$; and $f_{14}$ in $d_1, d_9, d_{20}, d_{32}$. While the search was designed to be as thorough as possible, some attributes may still have been missed—an issue that likely affects other researchers as well. Missing or hard-to-discover features can limit a dataset's applicability. For the purpose of this study, such gaps were recorded as NaN and imputed as zero. Documenting these omissions highlights opportunities for enriching dataset metadata to improve usability[3].

**Recommendations:** As shown in Table 4.5, $D_8$, $D_{17}$, $D_{18}$, and $D_{34}$ stand out for broad multi-task support. For instance, $D_8$ achieves $U_t^{(d)} \geq 1.25$ in $T_5$ and $T_6$; $D_{17}$ in $T_3$, $T_5$, and $T_7$; $D_{18}$ in $T_1$ and $T_{3-8}$, and $D_{34}$ in $T_{2-8}$. Based on Task-Lens findings, we propose five key principles for future dataset development:

1. Ensure high acoustic diversity and allow modular inclusion of underrepresented recording conditions.

[3]Some datasets are distributed under paid licenses, and missing details may be present in the licensed versions.

| **Dataset** | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ |
|---|---|---|---|---|---|---|---|---|
| $D_1$ | -0.65 | 0.31 | -0.13 | 0.15 | 0.19 | -0.61 | **0.54** | -0.40 |
| $D_2$ | **0.39** | -0.01 | -0.35 | -0.05 | -0.18 | -0.10 | -0.12 | 0.18 |
| $D_3$ | **0.95** | 0.61 | 0.72 | 0.87 | 0.77 | 0.90 | 0.73 | 0.27 |
| $D_4$ | **1.25** | 0.36 | 0.01 | 0.35 | 0.25 | 0.63 | 0.19 | 0.91 |
| $D_5$ | **-0.37** | -0.66 | -0.77 | -0.89 | -0.90 | -0.68 | -0.91 | -0.53 |
| $D_6$ | -0.15 | **-0.09** | -0.41 | -0.33 | -0.41 | -0.29 | -0.54 | -0.27 |
| $D_7$ | -0.34 | -0.09 | **0.18** | -0.16 | -0.07 | -0.04 | -0.20 | -0.11 |
| $D_8$ | 1.14 | 0.82 | 0.94 | 0.86 | 1.31 | **1.53** | 1.03 | 0.76 |
| $D_9$ | -0.26 | 0.20 | 0.16 | **0.61** | 0.44 | -0.27 | 0.86 | -0.05 |
| $D_{10}$ | **-0.80** | -1.48 | -1.41 | -1.66 | -1.34 | -1.02 | -1.31 | -1.35 |
| $D_{11}$ | -0.56 | -0.29 | **0.01** | -0.21 | -0.17 | -0.17 | -0.37 | -0.13 |
| $D_{12}$ | -3.78 | **-2.94** | -3.31 | -3.28 | -3.37 | -3.58 | -3.47 | -3.86 |
| $D_{13}$ | -0.66 | 0.00 | **0.30** | 0.27 | 0.27 | 0.11 | 0.15 | -0.32 |
| $D_{14}$ | -0.61 | 0.11 | 0.40 | 0.36 | **0.42** | 0.23 | 0.28 | -0.31 |
| $D_{15}$ | **0.67** | 0.48 | 0.25 | 0.48 | 0.62 | 0.53 | 0.61 | 0.36 |
| $D_{16}$ | -0.04 | **0.00** | -0.08 | -0.14 | -0.13 | -0.09 | -0.23 | -0.04 |
| $D_{17}$ | 1.01 | 1.25 | 1.28 | 1.20 | **1.36** | 1.13 | **1.36** | 0.94 |
| $D_{18}$ | 1.67 | 0.79 | 1.41 | 1.48 | 1.60 | **1.76** | 1.44 | 1.53 |
| $D_{19}$ | **-0.57** | -0.75 | -1.03 | -1.15 | -1.12 | -0.88 | -1.13 | -0.78 |
| $D_{20}$ | -0.80 | 0.01 | -0.19 | **0.35** | 0.00 | -0.46 | 0.24 | -0.41 |
| $D_{21}$ | 0.43 | 0.12 | **0.71** | 0.67 | 0.64 | 0.51 | 0.54 | 0.60 |
| $D_{22}$ | **-1.10** | -1.77 | -1.84 | -2.10 | -1.92 | -1.54 | -1.82 | -1.46 |
| $D_{23}$ | **0.03** | 0.00 | -0.19 | -0.15 | -0.21 | -0.05 | -0.24 | -0.11 |
| $D_{24}$ | **-0.24** | -0.28 | -0.54 | -0.44 | -0.56 | -0.41 | -0.64 | **-0.24** |
| $D_{25}$ | 0.76 | 0.15 | 0.32 | 0.44 | 0.55 | **1.01** | 0.65 | 0.68 |
| $D_{26}$ | -0.20 | -0.04 | 0.19 | 0.02 | -0.17 | 0.10 | -0.37 | **0.54** |
| $D_{27}$ | **0.66** | 0.36 | 0.01 | 0.23 | 0.16 | 0.19 | 0.19 | 0.44 |
| $D_{28}$ | 0.15 | -0.03 | -0.03 | -0.01 | -0.04 | 0.03 | -0.02 | **0.22** |
| $D_{29}$ | 0.71 | 0.42 | **1.93** | 0.90 | 0.72 | 0.56 | 0.82 | 0.97 |
| $D_{30}$ | 0.81 | 0.41 | 0.55 | 0.57 | 0.63 | 0.54 | 0.75 | **0.84** |
| $D_{31}$ | 1.04 | 0.12 | 0.71 | 0.79 | 0.73 | 0.95 | 0.54 | **1.07** |
| $D_{32}$ | -0.99 | -0.23 | -0.48 | -0.06 | -0.33 | -0.91 | **-0.08** | -0.78 |
| $D_{33}$ | **-0.55** | -1.34 | -0.96 | -1.40 | -1.17 | -0.91 | -1.01 | -0.74 |
| $D_{34}$ | 1.04 | **3.44** | 1.62 | 1.38 | 1.45 | 1.33 | 1.54 | 1.55 |

Table 4.5: Utility scores of each dataset ($D_i$) across all tasks ($T_1$–$T_8$). For each task, the top three datasets are shaded: light blue (highest), light red (second), and light green (third). For each dataset, its most relevant task is both bolded and underlined. Notably, $D_8$, $D_{17}$, $D_{18}$, and $D_{34}$ achieve scores above 0.8 on most tasks, indicating strong cross-task applicability. Higher scores correspond to greater task-specific utility. Task and dataset descriptions are provided in Section 4.3 and Table 4.1.

2. Recruit speakers from diverse demographic backgrounds to enhance ASR, speaker verification, and emotion recognition capabilities.

3. Provide multi-layered annotations—particularly normalized and phonetic transcripts—alongside utterance and speaker-level metadata for richer transcription and prosody modeling.

4. Adopt a standardized metadata schema covering speaker identity, recording environment, device type, noise conditions, and emotion labels to enable systematic filtering and targeted augmentation.

5. Publish per-task utility profiles with explicit thresholds (e.g., $U_t^{(d)} \geq 1.0$ for gen-

eralist readiness, $U_t^{(d)} < 0$ to indicate areas needing improvement) and maintain open, version-controlled repositories with mechanisms for community-driven updates to meet evolving multi-task requirements.

### 4.4.2 Task-Wise Data Requirement

Advancing robust speech models for Indian demographics requires a systematic mapping of available datasets for each task, along with the identification of persisting gaps. While ASR and TTS resources have grown substantially, dataset coverage across tasks remains uneven and often unclear to practitioners [65, 141]. Researchers often spend substantial time exploring repositories to verify the availability of data for tasks such as audio deepfake detection or speaker verification, only to find that datasets for underrepresented Indian languages are scarce or entirely absent [19]. These insights lead to our second research question: *Which tasks remain under-resourced for the Indian demographic population?*

**Setup:** We used the Task-Lens utility scores (Table 4.5) to select datasets with a threshold of $U_t^{(d)} \geq 1.0$ for each task. This selection yielded the most relevant datasets per task: $D_4$, $D_8$, $D_{17}$, $D_{18}$, $D_{31}$, and $D_{34}$ for $T_1$; $D_1$, $D_{17}$, and $D_{34}$ for $T_2$; $D_{17}$, $D_{18}$, $D_{29}$, and $D_{34}$ for $T_3$; $D_{17}$, $D_{18}$, and $D_{34}$ for $T_4$; $D_8$, $D_{17}$, $D_{18}$, and $D_{34}$ for $T_5$; $D_8$, $D_{17}$, $D_{18}$, $D_{25}$, and $D_{34}$ for $T_6$; $D_8$, $D_{17}$, $D_{18}$, and $D_{34}$ for $T_7$; $D_{18}$, $D_{31}$, and $D_{34}$ for $T_8$. We then analyzed the total speech duration of these high-utility datasets (Figure 4.3) to assess their coverage[4].

**Observations:** Figure 4.3 illustrates that tasks $T_1$, $T_5$, $T_6$, and $T_7$ achieve moderate dataset coverage. Tasks $T_3$ and $T_4$ have coverage of roughly 40–50 thousand minutes, indicating fair support that could be further enhanced. In contrast, tasks $T_2$ and $T_8$ each cover about 34 thousand minutes, revealing notable gaps in available synthetic speech and high-quality low-resource datasets.

**Recommendations:** Analysis indicates that tasks $T_2$ (ADD) and $T_8$ (TTS) need new datasets to address duration shortfalls. $T_2$ calls for synthetic dialectal speech encompassing numerous speakers, explicit gender labels, and a range of accents. $T_8$ requires

[4]For tasks with abundant data, a stricter threshold (e.g., $U_t^{(d)} \geq 1.5$) can isolate the strongest datasets, whereas for low-resource tasks, a relaxed threshold (e.g., $U_t^{(d)} \geq 0.25$) can expand the candidate pool. In this study, we set the threshold to 1.00.

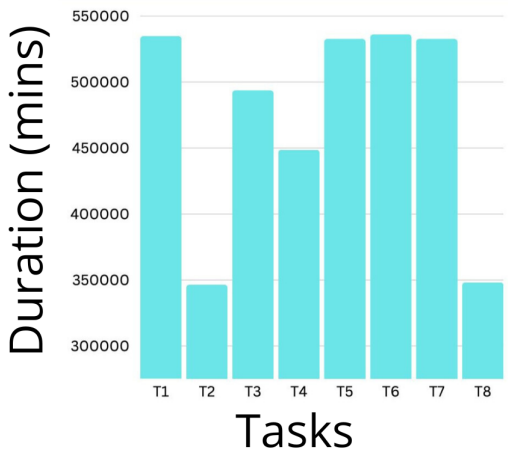



Figure 4.3: Total data duration (in minutes) available for each target downstream task.

expressive, monolingual recordings showcasing diverse voice styles and high-fidelity audio. All datasets should adhere to a consistent annotation framework and a rigorous quality-control process.

### 4.4.3 Linguistic Data Requirement

Researchers in the NLP community aim to build multilingual and inclusive speech technologies, yet most publicly accessible speech datasets remain English-centric. This scarcity of low-resource language datasets poses a significant barrier to advancing multilingual speech systems. In fact, only a small fraction of the world's $\approx$ 7,000 languages possess adequate resources for human language technologies [13]. This imbalance drives our third research question: *Which Indian languages have sufficient dataset coverage for each task, and where do key language-specific gaps remain?*

**Setup:** For each task, we selected all datasets with a utility score of $U_t^{(d)} \geq 1.0$. We then identified the languages represented in these datasets, resulting in a total of 26 languages. For each language, we computed the aggregate audio duration across all qualifying datasets. Table 4.6 reports the total speech durations (in hours) for datasets satisfying the strict $U_t^{(d)} \geq 1.0$ threshold, covering 26 languages and eight downstream tasks ($T_1$–$T_8$). Additionally, Table 4.7 lists the IDs of curated datasets containing that language, helping researchers locate relevant resources. Figure 4.4 visualizes the total speech duration for each language across all 33 datasets.

**Observations:** Table 4.6 and Figure 4.4 highlight pronounced imbalances in lan-

| **Language** | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ | $T_8$ |
|---|---|---|---|---|---|---|---|---|
| $L_1$: Assamese | 187.34 | **0.00** | 192.44 | 175.34 | 187.34 | 187.34 | 187.34 | 175.34 |
| $L_2$: Bengali | 394.16 | 267.23 | 396.22 | 379.22 | 391.22 | 391.22 | 391.22 | 263.42 |
| $L_3$: Bhojpuri | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| $L_4$: Bodo | 172.05 | 172.05 | **0.00** | 172.05 | 172.05 | 172.05 | 172.05 | 172.05 |
| $L_5$: Dogri | **70.68** | **0.00** | **70.68** | **70.68** | **70.68** | **70.68** | **70.68** | **70.68** |
| $L_6$: Garhwali | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| $L_7$: Gujarati | 347.16 | 326.22 | 335.16 | 335.16 | 347.16 | 347.16 | 347.16 | 205.86 |
| $L_8$: Hindi | 575.93 | 489.33 | 563.93 | 563.93 | 575.93 | 582.43 | 575.93 | 413.73 |
| $L_9$: Indian English | **9.60** | **19.82** | **0.00** | **0.00** | **0.00** | **32.03** | **0.00** | **9.60** |
| $L_{10}$: Kannada | 390.74 | 334.13 | 378.74 | 378.74 | 390.74 | 395.54 | 390.74 | 312.94 |
| $L_{11}$: Kashmiri | **64.99** | **0.00** | **64.99** | **64.99** | **64.99** | **64.99** | **64.99** | **64.99** |
| $L_{12}$: Konkani | **53.06** | **0.00** | **53.06** | **53.06** | **53.06** | **53.06** | **53.06** | **53.06** |
| $L_{13}$: Maithli | **81.77** | **0.00** | **81.77** | **81.77** | **81.77** | **81.77** | **81.77** | **81.77** |
| $L_{14}$: Malayalam | 336.83 | 242.26 | 324.83 | 324.83 | 336.83 | 341.53 | 336.83 | 177.53 |
| $L_{15}$: Manipuri | **23.99** | **0.00** | **23.99** | **23.99** | **23.99** | **23.99** | **23.99** | **23.99** |
| $L_{16}$: Marathi | 716.30 | 653.40 | 704.30 | 704.30 | 716.30 | 716.30 | 716.30 | 529.10 |
| $L_{17}$: Nepali | 120.67 | **0.00** | 105.87 | 105.87 | 117.87 | 117.87 | 117.87 | 105.87 |
| $L_{18}$: Odia | 286.19 | 203.24 | 274.19 | 274.19 | 286.19 | 286.19 | 286.19 | 162.59 |
| $L_{19}$: Punjabi | 606.55 | 519.61 | 594.55 | 594.55 | 606.55 | 606.55 | 606.55 | 457.65 |
| $L_{20}$: Rajasthani | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| $L_{21}$: Sanskrit | 1150.32 | 1114.57 | 1150.32 | 1150.32 | 1150.32 | 1150.32 | 1150.32 | 1034.82 |
| $L_{22}$: Santali | **76.37** | **0.00** | **76.37** | **76.37** | **76.37** | **76.37** | **76.37** | **76.37** |
| $L_{23}$: Sindhi | **22.48** | **0.00** | **10.48** | **10.48** | **22.48** | **22.48** | **22.48** | **10.48** |
| $L_{24}$: Tamil | 875.41 | 763.94 | 891.61 | 863.41 | 875.41 | 879.91 | 875.41 | 678.31 |
| $L_{25}$: Telugu | 658.36 | 509.96 | 646.36 | 646.36 | 658.36 | 662.66 | 658.36 | 491.46 |
| $L_{26}$: Urdu | 323.60 | 232.99 | 311.60 | 311.60 | 323.60 | 323.60 | 323.60 | 224.90 |

Table 4.6: The table presents speech durations (in hours) for datasets meeting the $U_t^{(d)} \geq 1.0$ criterion, spanning 26 languages and eight downstream tasks ($T_1$–$T_8$). For each task, the three longest durations are highlighted: light blue (highest), light red (second), and light green (third). Language–task pairs with less than 50 hours are bolded and underlined to emphasize scarcity. There is a critical shortage of datasets for $L_3$ (Bhojpuri), $L_7$ (Garhwali), and $L_{20}$ (Rajasthani) across all tasks.

guage coverage across tasks. High-resource languages like $L_{24}$ (Tamil) and $L_{25}$ (Telugu) lead in most tasks, supported by large multi-speaker datasets contributing hundreds of hours of audio. In contrast, $L_3$ (Bhojpuri), $L_6$ (Garhwali), and $L_{20}$ (Rajasthani) show critical gaps, with urgent need for datasets across all tasks. Meanwhile, $L_7$ (Gujarati), $L_{14}$ (Malayalam), and $L_{26}$ (Urdu) display consistent but relatively modest coverage. Overall, the results reveal that essential speech data is heavily concentrated in a few languages, leaving many Indian languages significantly underrepresented.

**Insights:** Our analysis highlights the unique value each dataset brings to the table. $D_5$ strengthens Bengali-specific modeling, improving linguistic accuracy; $D_6$ and $D_{24}$ capture speech from numerous speakers in low-resource languages, aiding bias detection and noise-resilience research; $D_{10}$ delivers broad multilingual coverage, making it valuable for cross-language acoustic studies; $D_{16}$, with its rich feature diversity, is

| Language | Dataset |
|---|---|
| $L_1$: Assamese | $D_{i\in\{2,3,8,13-15,18,29\}}$ |
| $L_2$: Bengali | $D_{i\in\{3-6,8,13-18,29,34\}}$ |
| $L_3$: Bhojpuri | $D_{i\in\{9,13\}}$ |
| $L_4$: Bodo | $D_{i\in\{15,18\}}$ |
| $L_5$: Dogri | $D_{i\in\{15,18\}}$ |
| $L_6$: Garhwali | $D_{i\in\{9,13\}}$ |
| $L_7$: Gujarati | $D_{i\in\{3,8,13-15,17,18,22,23,27,34\}}$ |
| $L_8$: Hindi | $D_{i\in\{3,8,9,11,12,14-18,21,23-25,31,34\}}$ |
| $L_9$: Indian English | $D_{i\in\{1,12,25,26,31\}}$ |
| $L_{10}$: Kannada | $D_{i\in\{3,8,13-15,17,18,25,27,34\}}$ |
| $L_{11}$: Kashmiri | $D_{i\in\{14,18,19\}}$ |
| $L_{12}$: Konkani | $D_{i\in\{14,15,18\}}$ |
| $L_{13}$: Maithli | $D_{i\in\{15,18\}}$ |
| $L_{14}$: Malayalam | $D_{i\in\{3,8,14-18,25,27,30,34\}}$ |
| $L_{15}$: Manipuri | $D_{i\in\{14,15,18\}}$ |
| $L_{16}$: Marathi | $D_{i\in\{3,7,8,13-15,17,18,23,27,34\}}$ |
| $L_{17}$: Nepali | $D_{i\in\{3,4,6,8,14,15,18\}}$ |
| $L_{18}$: Odia | $D_{i\in\{3,8,14,15,17,18,23,34\}}$ |
| $L_{19}$: Punjabi | $D_{i\in\{3,8,13,14,15,17,18,34\}}$ |
| $L_{20}$: Rajasthani | $D_{i\in\{13,15,16,28\}}$ |
| $L_{21}$: Sanskrit | $D_{i\in\{15,17,18,33,34\}}$ |
| $L_{22}$: Santali | $D_{i\in\{18\}}$ |
| $L_{23}$: Sindhi | $D_{i\in\{8,15,18\}}$ |
| $L_{24}$: Tamil | $D_{i\in\{3,8,10,12-15,17,18,22,23,25,27,29,34\}}$ |
| $L_{25}$: Telugu | $D_{i\in\{3,8,14,15,17,18,22,23,25,27,34\}}$ |
| $L_{26}$: Urdu | $D_{i\in\{3,8,13,17,18,32,34\}}$ |

Table 4.7: Mapping between the 26 Indian languages ($L_1$–$L_{26}$) considered in this study and the curated datasets ($D_i$) in which they appear. Dataset IDs correspond to the numbering used in Table 4.1.

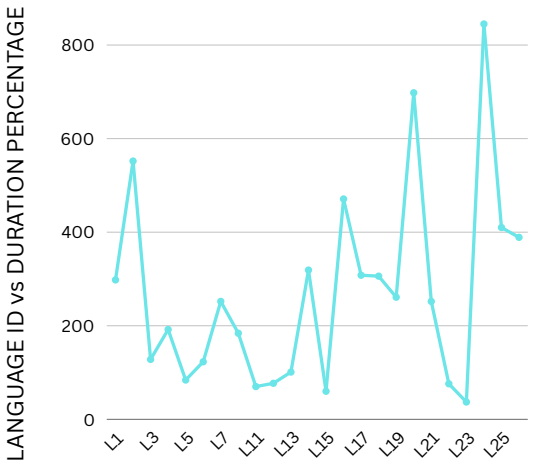



Figure 4.4: Task-wise cumulative speech duration for each Indian language ($L_1$–$L_{26}$) across all 33 datasets. $L_8$ (Hindi) and $L_9$ (Indian English) contain 2,487 and 15,953 hours of data, respectively, and were omitted from the figure to prevent their sheer size from overshadowing differences among other languages. High-resource languages such as $L_2$ (Bengali) and $L_{24}$ (Tamil) record the largest durations, while low-resource languages like $L_3$ (Bhojpuri), $L_7$ (Garhwali), and $L_{20}$ (Rajasthani) appear almost entirely absent, aside from limited representation in emotion-related tasks.

a key TTS resource for Indian languages; $D_{19}$ and $D_{32}$ address critical gaps in Kashmiri and Urdu, respectively, fostering broader language inclusion; $D_{22}$ spans multiple Indian languages with a large, consistently recorded speaker base; and $D_{33}$, equipped with detailed speaker background metadata, supports high-quality Sanskrit TTS and enables demographic-sensitive modeling. This breadth—from small, specialized sets to expansive multilingual corpora—empowers researchers to optimize dataset utility, create fair and representative acoustic models, mitigate resource shortages, and push forward speech technology for all languages.

**Recommendations:** Future dataset development should prioritize underrepresented languages with targeted, task-specific data collection. For languages without foundational ASR resources, even modest pilot corpora can jump-start fine-tuning of multilingual models. Community-led efforts should be fostered to capture spontaneous, conversational, and code-switched speech—particularly for $L_3$ (Bhojpuri), $L_7$ (Garhwali), and $L_{20}$ (Rajasthani). Existing multilingual datasets can be enriched with annotated emotional content and adversarial samples to expand task coverage without requiring complete re-collection. Lastly, creating shared benchmarks and open version-controlled repositories for each Indian language will enable incremental dataset growth, ensure consistent cross-language evaluation, and advance inclusive, domain-resilient speech technologies.

## 4.5 Conclusions and Future Work

This chapter presented Task-Lens, a novel framework for cross-task, utility-based profiling of speech datasets. The approach comprises four key stages: dataset discovery, dataset filtering, feature extraction, and utility mapping, ultimately producing a task-specific utility score for each dataset. Task-Lens was applied to evaluate 34 Indian speech datasets spanning 26 languages and totaling 74,745 hours of audio. Sixteen descriptive features—covering audio characteristics, metadata and licensing, speaker demographics, speech style, and paralinguistic aspects—were used for utility score computation. Scores were generated for eight target tasks, enabling the assessment of four core research questions: (1) the degree to which each dataset supports individual tasks, (2) possible improvements for enhancing cross-task applicability, (3) identifi-

cation of tasks with inadequate data support, and (4) evaluation of language-specific coverage and existing gaps. Results showed that FLEURS [34], INDICSUPERB [65], INDICVOICES-R [141], and IndicSynth attained the top rankings. Many other datasets lacked sufficient metadata diversity and scale. While automatic speech recognition and multilingual text-to-speech are relatively well-supported, tasks such as audio deepfake detection, emotion recognition, and certain TTS resources remain significantly underdeveloped. Language-wise, Bengali and Tamil offer moderate coverage, Bhojpuri, Rajasthani, and Garhwali are scarcely represented, and Indian English dominates in terms of total duration.

The work also outlines several directions for future research. By providing clear utility scores and highlighting dataset limitations, Task-Lens offers researchers a practical means to select suitable datasets, thereby allowing greater focus on model development rather than dataset curation. It facilitates rapid dataset discovery by identifying optimal resources for specific tasks. Moreover, the framework highlights under-resourced languages and tasks, guiding the creation of new datasets where they are most needed. Extending Task-Lens to include additional tasks, languages and dialects—within India and internationally—could further enhance its scope and support the development of inclusive and domain-robust speech technologies.

## 4.6 Limitations

While Task-Lens offers broad applicability and practical value, several limitations should be noted:

1. **Language and Task Scope:** The current evaluation encompasses eight speech tasks, 26 languages, andgn 34 publicly available Indian speech datasets. Proprietary, restricted-access, or emerging datasets—such as those for code-switching ASR—are excluded. Future work can incorporate both proprietary and community-generated datasets to expand language and task coverage.

2. **Feature Weight Assignment:** The calculation of utility scores is based on manually assigned weights, reflecting each task's functional requirements and the empirical importance of features reported in prior research. This introduces the possibility of bias. These scores are not intended as direct predictors of model performance, but rather as a structured means of identifying datasets with potential suitability for specific tasks. By publishing these rankings, we aim to encourage further community-led benchmarking and comparative evaluations—particularly

for low-resource datasets—to examine how closely utility-based rankings align with actual performance improvements.

3. **Static Nature of Evaluation:** Task-Lens currently operates as a one-time evaluation and does not automatically update with the release of new datasets or changes in standards. This limitation could be addressed by implementing an automated pipeline with periodic benchmarking to maintain up-to-date utility scores.

Despite these constraints, Task-Lens provides the NLP community with a transparent and practical foundation for dataset selection, a clear roadmap for extending evaluations to additional languages and tasks, and actionable insights into resource gaps that can guide targeted data creation and sharing.

## 4.7 Ethical Considerations

The adoption of Task-Lens can promote systematic, cross-task profiling of speech datasets across multiple languages and tasks. However, researchers should adhere to ethical obligations—particularly in verifying dataset licenses, reviewing README documentation, and ensuring compliance with usage restrictions before using any dataset in research or commercial applications.

# CHAPTER 5

# Mitigating Linguistic Bias in Speaker Verification

Linguistic bias in Deep Neural Network (DNN)-based Natural Language Processing (NLP) systems is a well-recognized challenge, and its implications become more severe in security-sensitive applications such as audio deepfake detection (ADD) and speaker verification (SV), where fairness and reliability are critical. To facilitate bias mitigation research in ADD and anti-spoofing, Chapter 2 introduced the 4000-hour IndicSynth dataset covering 12 Indian languages. The chapter highlighted the vulnerability of SV models against multilingual audio spoofing attacks. Furthermore, to facilitate bias mitigation research across more speech-based tasks, Chapter 4 presented a systematic cross-task profiling of 34 Indian speech datasets covering 26 languages across eight downstream tasks.

This chapter addresses the problem of linguistic bias in SV. SV systems aim to determine whether two speech recordings belong to the same speaker, and such systems must remain accessible to speakers of diverse languages. Ideally, an SV model trained on one language should generalize effectively to others. However, DNN-based approaches often exhibit language dependency. Prior studies have investigated domain adaptation strategies that fine-tune pre-trained models for out-of-domain languages, but performing language-specific adaptation for every case is computationally expensive and reduces practical scalability.

To address this limitation, this chapter[1] introduces a cost-efficient approach that integrates a lightweight embedding into existing SV frameworks to mitigate linguistic bias without requiring language-specific adaptation. The method is grounded in the hypothesis that attentive frames can help generate language-agnostic embeddings. To validate this hypothesis, we design two frame-attentive networks, *FAtNet-v1* and *FAtNet-v2*, and evaluate their integration with standard baselines across twelve global languages.

[1]This chapter presents the following paper:
Divya V Sharma and Arun Balaji Buduru. 2022. FAtNet: 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.

Experimental findings demonstrate that frame-attentive embeddings effectively reduce linguistic bias in a cost-efficient manner, thereby improving the usability of SV systems in multilingual settings.

## 5.1 Introduction

Mitigating linguistic bias in Deep Neural Network (DNN)-based models is a critical challenge in Natural Language Processing (NLP). This issue becomes particularly significant in security applications, such as speaker verification (SV) systems, where fairness and reliability are essential. Speaker verification systems are biometric authentication tools that use speech signals to verify a speaker's identity, leveraging the fact that each individual has unique vocal characteristics, as depicted in the Figure 5.1 [53]. These systems have real-world applications in areas including e-commerce, forensics, law enforcement, business, and access control [53]. SV systems can be either text-dependent or text-independent [53], with text-independent systems being more user-friendly, as they authenticate speakers without restrictions on the spoken content.

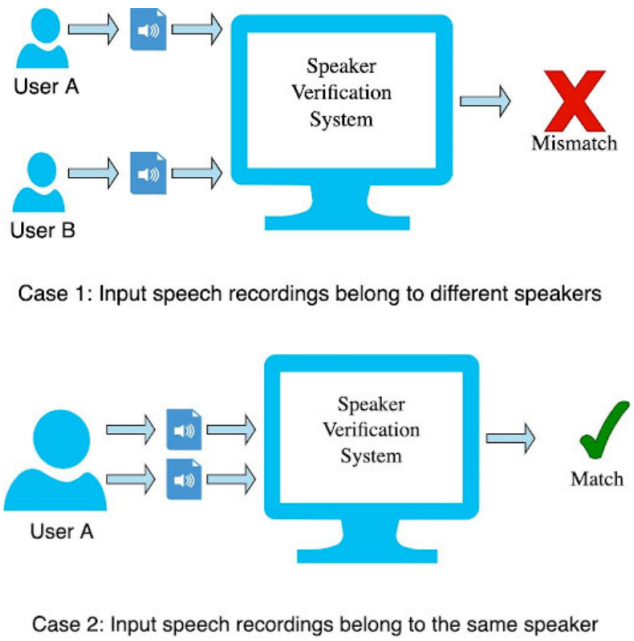



Figure 5.1: Speaker Verification System

Despite their utility, speaker verification models often exhibit language dependency [7]. This arises because robust SV systems rely on memory mechanisms to analyze sequential speech data and capture relevant discriminatory information. Such memory components, while useful for remembering past information and predicting future

frames, can inadvertently encode linguistic content into the embeddings [144]. Consequently, embeddings generated for speaker verification may contain language-specific details.

Language-dependent SV models tend to perform well when the test data matches the training language. However, their performance deteriorates on test sets comprising speech in different languages. Most publicly available speech datasets are in English, making it challenging to obtain labeled datasets for many low-resource languages. Prior works often employ domain adaptation to improve SV performance for a limited set of languages [137, 183, 24], but fine-tuning pre-trained models individually for each language is computationally expensive. Moreover, studies indicate that linguistic information in embeddings increases with the temporal scope of the representations [30].

Our approach is motivated by the hypothesis that frame-level features, due to their low temporal scope, contain minimal linguistic information and can therefore be used to generate language-agnostic embeddings. Further, an intelligent selection of frame-level features may enhance model generalizability to out-of-domain data. We aim to address language dependency in text-independent SV systems in a cost-efficient manner, without the overhead of domain adaptation. To this end, this chapter proposes integrating a lightweight embedding into existing SV frameworks to improve their cross-lingual generalizability.

To validate this hypothesis, we introduce two variants of frame-attentive networks, FAtNet-v1 and FAtNet-v2. These models take a pair of speech recordings as input and determine whether they belong to the same speaker, even if the speakers are previously unseen. We qualitatively evaluate the generalization ability of our models against two strong baselines on four publicly available datasets. Quantitative experiments are conducted on twelve languages to assess the integration of FAtNet embeddings with baseline models on out-of-domain test sets, without any domain adaptation.

**Summary of Chapter Contributions:**

1. Explore cognitive-inspired concepts such as attention mechanisms, residual connections for memory retention, and learning parameters to develop language-agnostic embeddings.
2. Formally test the theoretical hypothesis by introducing two frame-attentive architectures: **FAtNet-v1** and **FAtNet-v2**.
3. Conduct both qualitative and quantitative evaluations across twelve global lan-

guages, benchmarking against two strong baselines and utilizing four publicly available datasets.

## 5.2 Background and Motivation

### 5.2.1 Language dependency in speaker verification

Modern speaker recognition systems primarily rely on deep neural networks (DNNs) for feature extraction and identification tasks [53, 92, 73, 71, 114, 153, 113, 52, 200, 48]. However, most DNN-based models exhibit strong language dependency [120], which can limit their usability across geographically and linguistically diverse populations. Obtaining labeled datasets for low-resource languages remains a significant challenge [17]. Models that perform consistently across multilingual datasets can potentially benefit other applications, such as code-switching, by facilitating information transfer [12]. Considering that there are roughly 7,000 languages spoken worldwide [59], addressing linguistic bias in NLP models is crucial for improving their global applicability [59].

### 5.2.2 Recent approaches

Transfer learning has been explored as a strategy to mitigate domain mismatches, but acquiring sufficient labeled data for low-resource languages continues to be difficult [17]. Adversarial domain adaptation methods have been proposed to tackle cross-lingual speaker verification tasks [137, 183, 24, 17], but these approaches typically improve performance only for a limited set of languages and introduce the additional overhead of domain adaptation.

The work most closely related to our objective is [27], where the authors attempt to reduce linguistic bias in speaker verification without domain adaptation. They demonstrate that training on multiple languages can enhance generalization to out-of-domain languages. However, their approach requires extremely large datasets—over 196,000 speakers and 20 million utterances—which is computationally expensive and often impractical. Additionally, their method combines text-dependent and text-independent SV systems. In contrast, our approach focuses on integrating a lightweight embedding into

existing text-independent SV models to mitigate linguistic bias efficiently.

### 5.2.3 Linguistic content in frames

Speech signals are non-stationary, so they are typically divided into short frames, within which the signal is assumed to be approximately stationary [98]. The temporal span of each frame is usually just a few milliseconds. Studies in neural phonology have shown that representational similarity analysis (RSA) applied to local frame-level features yields weaker correlations between phonemes and neural activation patterns [30], suggesting that individual frames carry limited linguistic information.

### 5.2.4 Theoretical hypothesis

We hypothesize that utterance-level embeddings capture more linguistic content than frame-level embeddings, making frame-level representations more robust to language variation. Leveraging frame-level features allows the model to focus on speaker-specific discriminatory information while minimizing the influence of language content. Previous studies indicate that certain frames contribute more significantly to the final encoded representation [55], and attention mechanisms have been widely used in state-of-the-art SV models to selectively emphasize these critical frames [204, 119]. In this chapter, we explore how attention can be used to intelligently select frame-level features for creating language-agnostic embeddings.

## 5.3 Proposed Approach

To evaluate our theoretical hypothesis, we introduce two variants of Frame-Attentive Networks: FAtNet-v1 and FAtNet-v2.

As depicted in Figures 5.2 and 5.3, both FAtNet variants share similar time-delay neural network (TDNN) paths, with the following specifics: The models take as input a pair of Mel-frequency cepstral coefficients (MFCCs) for speaker verification [24, 204, 80]. MFCCs refine the features to align with human auditory perception [96]. Let $d$ denote the dimension of the input MFCCs, and $l_1$ and $l_2$ represent the number of

frames in the given pair for speaker verification, which may differ due to variable speech durations. The models are trained on 3-second speech segments [113], generating 80-dimensional MFCCs of shape (94, 80) as input.

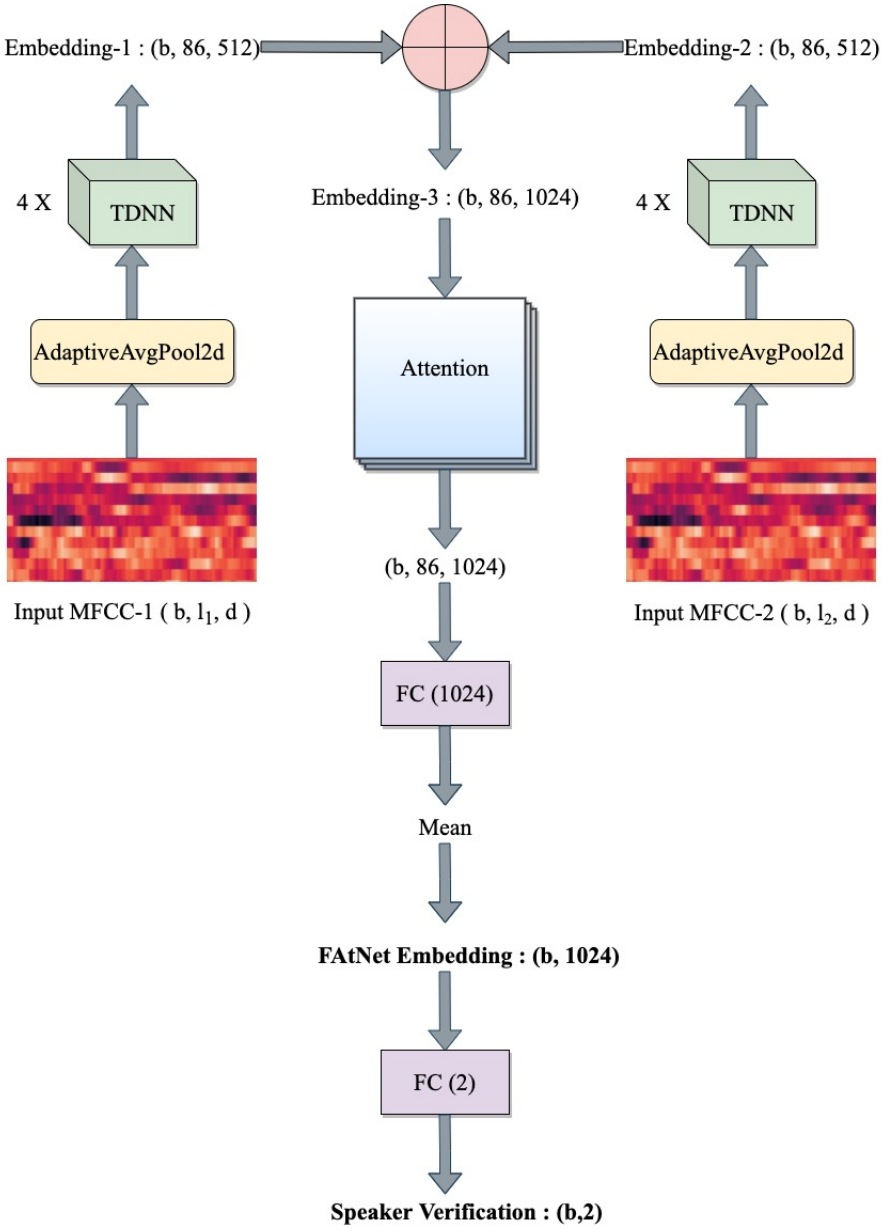



Figure 5.2: Architecture diagram for FAtNet-v1.

As shown in Figures 5.2 and 5.3, each input MFCC is passed through an AdaptiveAvgPool2d layer, producing features of shape $(b, 94, 80)$, where $b$ is the batch size [197]. This design allows the model to handle variable-duration speech recordings during testing without requiring specialized augmentation strategies. Additionally, it simplifies integration of FAtNet embeddings with other speaker verification models and improves overall model usability.

The next step involves computing frame-level features for further analysis. By lever-

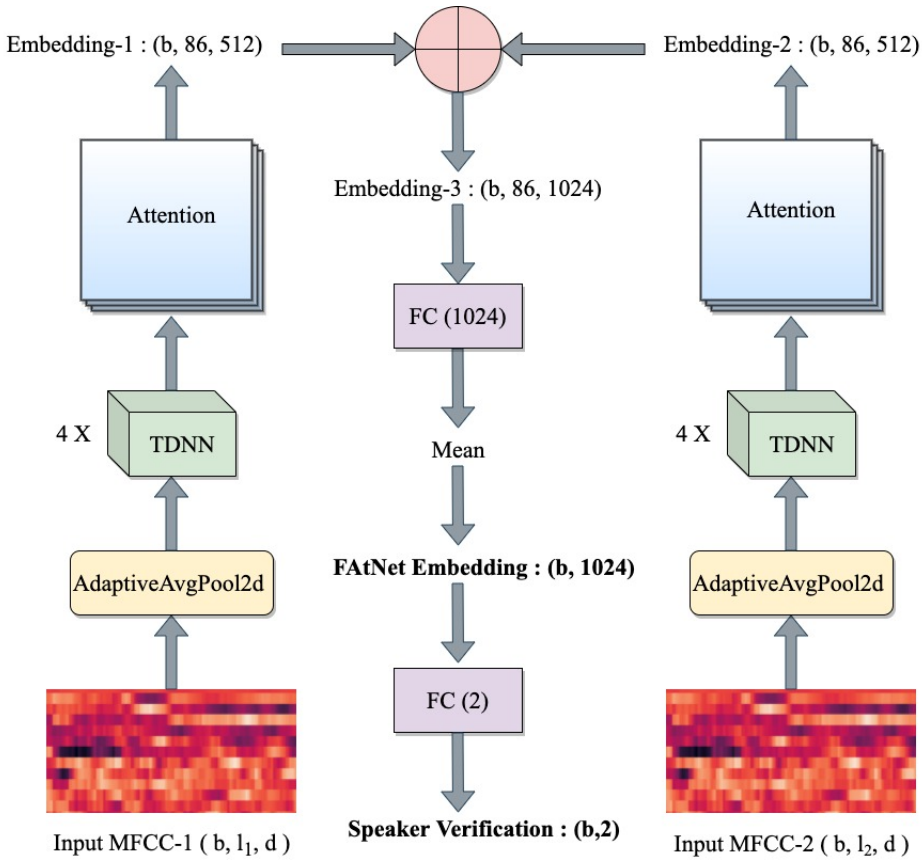



Figure 5.3: Architecture diagram for FAtNet-v2.

aging the abstract representation of frame-level features, the linguistic content in the final embedding can be minimized. We employ four stacked TDNN layers to extract these frame-level features, as illustrated in Table 5.1 [173]. Since the task involves determining whether a pair of speech recordings belong to the same speaker, two parallel TDNN paths are used for the input audio clips.

| **Hyper-parameter** | **#1** | **#2** | **#3** | **#4** |
|---|---|---|---|---|
| Input dimension | 80 | 512 | 512 | 512 |
| Output dimension | 512 | 512 | 512 | 512 |
| Context-size | 3 | 5 | 3 | 1 |
| Batch-norm | False | False | True | True |

Table 5.1: Hyper-parameter detail for the stacked TDNN layers in FAtNet models.

**FAtNet-v1:** For this variant, the frame-level features from both input speech recordings are concatenated, as illustrated in Figure 5.2. Batch normalization is applied to the concatenated features, which are then passed through an eight-head frame-level attention block. The attention mechanism assigns higher weights to the most informative features within each frame.

**FAtNet-v2:** In this variant, the frame-level features of each input recording are

passed through separate four-head attention blocks, resulting in embeddings 1 and 2 (Figure 5.3). Within each frame, the attention block emphasizes relevant features. The outputs from both attention blocks are then concatenated to produce embedding 3.

**Attention Mechanism:** Attention mechanisms are widely used in state-of-the-art speaker verification models [204, 179]. Our FAtNet attention design draws inspiration from [172, 107]. Frame-level features are processed through a multi-head residual self-attention block (Figures 5.2 and 5.3). The input to the attention block is a tensor of shape $(b, l, d)$, where $b$ represents the batch size, $l$ is the number of frames, and $d$ is the number of features per frame. Let $dv$ denote the dimension of the linear projection space and $nv$ the number of attention heads. The same tensor is used as the query, key, and value in the self-attention computation.

Each of the query, key, and value tensors is first passed through separate fully connected layers with $dv \times nv$ output units, followed by a $ReLU$ activation, resulting in modified tensors $Q$, $K$, and $V$. After appropriate reshaping, these tensors have dimensions $(b, l, nv, dv)$. For each example $i$, the attention block performs the following computation using $Q$, $K$, and $V$:

1. $K_{\text{i}}^{\text{permute}} := K_{\text{i}}.permute(0, 2, 1)$
2. $prob_{\text{i}} := Q_{\text{i}} * K_{\text{i}}^{\text{permute}}$
3. $prob_{\text{i}}^{\text{scaled}} := \frac{prod_{\text{i}}}{\sqrt{dv}}$
4. $weights_{\text{i}}^{\text{attn}}$:=Softmax($prod_{\text{i}}^{\text{scaled}}$, dim = -1)
5. $rprod_{\text{i}} := weights_{\text{i}}^{\text{attn}} * V_{\text{i}}$

We also incorporate a residual connection that functions as a memory mechanism, combining the original set of frame-level features with $rprod_{\text{i}}$. Specifically, the output of the residual attention block is computed by adding $rprod_{\text{i}}$ to the initial query tensor and passing the result through a fully connected layer with $d_{\text{out}}$ neurons, followed by a $ReLU$ activation.

The subsequent layers are consistent across both FAtNet-v1 and FAtNet-v2. The attentive frame features are first batch normalized and then processed through a fully connected layer for fine-grained analysis. A leaky-$ReLU$ activation is applied, followed by $L_2$-normalization. The frame-level features are then aggregated by computing their mean and passed through a final fully connected layer with two output units for speaker verification.

## 5.4 Experimental Setup

### 5.4.1 Datasets

#### Training datasets

Separate FAtNet-v1 models were trained on the publicly available VoxCeleb-1[2] (Vox-1 dev) and VoxCeleb-2 development sets (Vox-2 dev)[3] [114, 31], while FAtNet-v2 was trained on the VoxCeleb-2 development set [200]. VoxCeleb-1 contains recordings from 1,251 speakers, including 799 from the USA and 215 from the UK, where English predominates. Its development set comprises utterances from 1,211 speakers, with the test set including 40 speakers. The VoxCeleb-2 development set includes 5,994 speakers. We adhered to the dev-test splits provided in [114, 31]. Both datasets primarily consist of English speech recordings [25].

#### Test datasets

Experiments were performed using trial pairs from several publicly available datasets: the VoxCeleb-1[4] test set (predominantly English) [114], LibriSpeech[5] test set (English) [123], Aishell-1[6] test set (Mandarin) [18], and Voxforge[7] test set (covering ten languages: Bulgarian, Dutch, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish) [176]. For LibriSpeech, Aishell-1, and Voxforge, trial pairs were randomly generated from the respective datasets. The VoxCeleb-1, LibriSpeech, Aishell-1, and Voxforge test sets contained 37720, 47402, 23800, and 51856 trial pairs, respectively. Since most publicly available speech datasets, including VoxCeleb (used for model training), primarily consist of English recordings, we focus on evaluating the effectiveness of our approach on non-English test sets without applying domain adaptation.

[2]VoxCeleb-1: `https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html`
[3]VoxCeleb-2: `https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html`
[4]VoxCeleb-1: `https://www.robots.ox.ac.uk/~vgg/data/voxceleb/meta/veri_test.txt`
[5]LibriSpeech: `https://www.openslr.org/12`
[6]Aishell-1: `https://www.openslr.org/33/`
[7]Voxforge: `http://www.voxforge.org/`

### 5.4.2 Training Setup

We preprocessed the speech recordings by removing silent segments and dividing them into 3-second clips. Acoustic features were extracted as 80-dimensional MFCCs using the Librosa library [102]. For training, we randomly generated separate datasets consisting of 525,000 and 2,388,000 trial pairs for models trained on VoxCeleb-1 dev and VoxCeleb-2 dev, respectively. Each training set contained an equal number of positive and negative trial pairs, and the examples were shuffled at the start of each epoch. The batch size was set to 128.

The proposed models were trained under joint supervision using softmax loss and center loss. While softmax loss increases inter-speaker separability, center loss reduces intra-speaker variations [92]. Adam optimizer was used for FAtNet, and RMSProp was used for center loss [92, 125]. For VoxCeleb-1 dev, the learning rates were 0.005 for Adam and 0.2 for RMSProp, with step-wise decay applied every ten epochs using a gamma of 0.5 for Adam and 0.3 for RMSProp. For VoxCeleb-2 dev, a lower learning rate of 0.0005 was applied to both Adam and RMSProp due to the larger number of steps per epoch. The total loss was calculated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{softmax}} + 0.01\,\mathcal{L}_{\text{center}}$$

Training was performed on a GeForce GTX 1080 GPU.

### 5.4.3 Baselines

Experiments were conducted using two publicly available baseline models: RawNet-2[8] [73] and VGG-M[9] [114].

#### RawNet-2

RawNet-2 is an enhanced version of RawNet [71] that directly processes raw audio waveforms to extract speaker embeddings. The model is pre-trained on the VoxCeleb-2

[8]Pre-trained RawNet-2 model available at `https://github.com/Jungjee/RawNet`

[9]Pre-trained VGG-M model available at `https://github.com/Derpimort/VGGVox-PyTorch`

dataset for speaker identification, producing 1024-dimensional embeddings [73]. Each speech recording in a trial pair is input to the model separately, generating two 1024-dimensional embeddings—one for each audio. A cosine similarity score is then computed between these embeddings to perform the speaker verification task.

#### VGG-M

The VGG-M model was trained on the full VoxCeleb-1 dataset for speaker identification [114] and produces a 4096-dimensional discriminative embedding. We utilized this pre-trained model to build a siamese network for speaker verification, which was fine-tuned on the VoxCeleb-1 development set. Each speech recording in a trial pair is passed through the VGG-M models with frozen weights. The resulting 4096-dimensional embeddings are concatenated to form a single 8,192-dimensional embedding. After applying batch normalization, this embedding is processed through a fully connected layer with 512 units and a ReLU activation. Finally, following $L_2$-normalization, the 512-dimensional VGG embedding is passed to another fully connected layer with two units for speaker verification.

### 5.4.4 Input strategy

For simplicity, the input features are fed into the model without any test-time augmentation. The adaptive average pooling layer in FAtNet addresses the variability in speech duration. Unlike a siamese network, FAtNet learns the weights of the two TDNN paths independently. Features from each audio clip in the trial pair are passed through both TDNN paths. We then compute their mean as follows: let $mfcc_1$ and $mfcc_2$ represent the MFCC features for the trial pair clips.

**FAtNet-v1**:

1. $\text{prob}_1$ := model($\text{mfcc}_1$, $\text{mfcc}_2$)
2. $\text{prob}_2$ := model($\text{mfcc}_2$, $\text{mfcc}_1$)
3. $\text{prob}_{\text{final}}$ := mean($\text{prob}_1$, $\text{prob}_2$)

**FAtNet-v2**:

1. $\text{emb}_{1a}$,$\text{emb}_{2a}$ := model($\text{mfcc}_1$, $\text{mfcc}_2$)

2. $emb_{2b}$,$emb_{1b}$ := model($mfcc_2$, $mfcc_1$)
3. $emb_1$ := mean($emb_{1a}$, $emb_{1b}$)
4. $emb_2$ := mean($emb_{2a}$, $emb_{2b}$)
5. $prob_{final}$ := CosineSimilarity($emb_1$, $emb_2$)

### 5.4.5 Evaluation Metric

Equal Error Rate (EER) is a widely used metric for evaluating biometric systems [53]. Accordingly, we assess the effectiveness of our approach using EER, where a lower value corresponds to better performance.

## 5.5 Experiments and Results

### 5.5.1 Experimental validation of hypothesis

Our approach is designed to mitigate linguistic bias in existing speaker verification systems by incorporating a language-agnostic embedding. To empirically validate this hypothesis, we examine the impact of integrating FAtNet embeddings with baseline models on out-of-domain test sets.

Consider a trial pair $(clip1.wav, clip2.wav)$ represented by MFCCs $(mfcc_1, mfcc_2)$ and spectrograms $(spec_1, spec_2)$.

**VGG-M $\oplus$ FAtNet-v1:** In this setup, the MFCCs $(mfcc_1, mfcc_2)$ are processed through FAtNet-v1 to generate 1024-dimensional embeddings. Simultaneously, the spectrograms $(spec1, spec2)$ are passed through the VGG-M siamese baseline to obtain 512-dimensional embeddings. These embeddings are then concatenated and fed into a fully connected layer with 1024 neurons. After applying $ReLU$ activation and $L_2$ normalization, the features pass through a final fully connected layer with 2 units for speaker verification. The last two fully connected layers are fine-tuned using the VoxCeleb training set.

**RawNet-2 $\oplus$ FAtNet-v2:** For this integration, the MFCCs $(mfcc_1, mfcc_2)$ are fed through FAtNet-v2 to produce 512-dimensional embeddings for each recording, following steps 3 and 4 of the input strategy described in Section 5.4.4. Concurrently, the

| Model | FAtNet Train set | Test Set | EER (%) | Rel. Imp. (%) |
|---|---|---|---|---|
| VGG-M | - | Voxforge | 9.190 | - |
| VGG-M $\oplus$ **FAtNet-v1** | Vox-1 dev | Voxforge | 7.665 | **+16.594%** |
| VGG-M $\oplus$ **FAtNet-v1** | Vox-2 dev | Voxforge | 7.618 | **+17.106%** |
| VGG-M | - | Aishell-1 | 9.999 | - |
| VGG-M $\oplus$ **FAtNet-v1** | Vox-1 dev | Aishell-1 | 9.139 | **+8.601%** |
| VGG-M $\oplus$ **FAtNet-v1** | Vox-2 dev | Aishell-1 | 6.866 | **+31.333%** |
| RawNet-2 | - | Voxforge | 7.012 | - |
| RawNet-2 $\oplus$ **FAtNet-v2** | Vox-2 dev | Voxforge | 5.341 | **+23.831%** |
| RawNet-2 | - | Aishell-1 | 6.202 | - |
| RawNet-2 $\oplus$ **FAtNet-v2** | Vox-2 dev | Aishell-1 | 3.832 | **+38.213%** |

Table 5.2: Table illustrating the relative performance gains of VGG-M and RawNet-2 baselines following integration with FAtNet embeddings.

recordings are processed through the RawNet-2 baseline to obtain 1024-dimensional embeddings. For each trial pair, the embeddings from FAtNet-v2 and RawNet-2 are concatenated for both recordings, and cosine similarity is computed for speaker verification.

**Observations:** As shown in Table 5.2, integrating FAtNet embeddings leads to notable improvements in baseline performance on out-of-domain test sets. This indicates that, with minimal additional overhead, FAtNet embeddings can enhance the effectiveness of these baselines on unseen languages without requiring domain adaptation.

### 5.5.2 Language-specific analysis

For a more comprehensive evaluation of the findings from the previous experiment, we generated separate test sets for 11 languages using the Voxforge dataset. The Bulgarian test set contains 3,110 trial pairs, while each of the remaining test sets comprises 20,000 trial pairs. Figures 5.4 and 5.5 demonstrate that integrating the baselines with the proposed FAtNet embeddings consistently lowers the equal error rate. This further confirms that FAtNet embeddings help reduce the language dependency of baseline models and enhance their generalizability on out-of-domain test sets. Notably, an absolute improvement of 2.64% was observed on the Dutch (Non-English) test set after combining FAtNet-v1 with VGG-M, and a 3.65% absolute improvement was observed on the Portuguese (Non-English) set after integrating RawNet-2 with FAtNet-v2. Hence, the largest absolute gains were seen on Non-English test sets, specifically Dutch and

Portuguese.

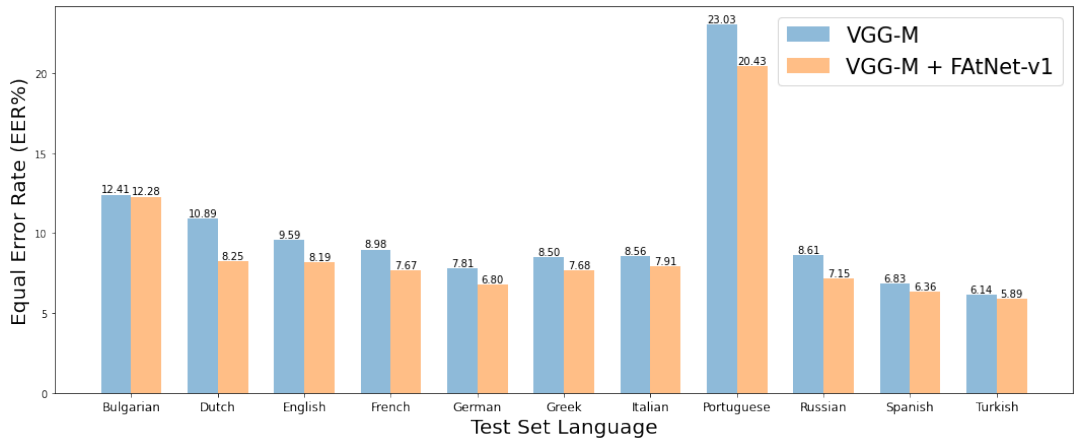



Figure 5.4: Figure illustrating that combining VGG-M with FAtNet-v1 consistently lowers the EER across test sets containing speech in various languages.

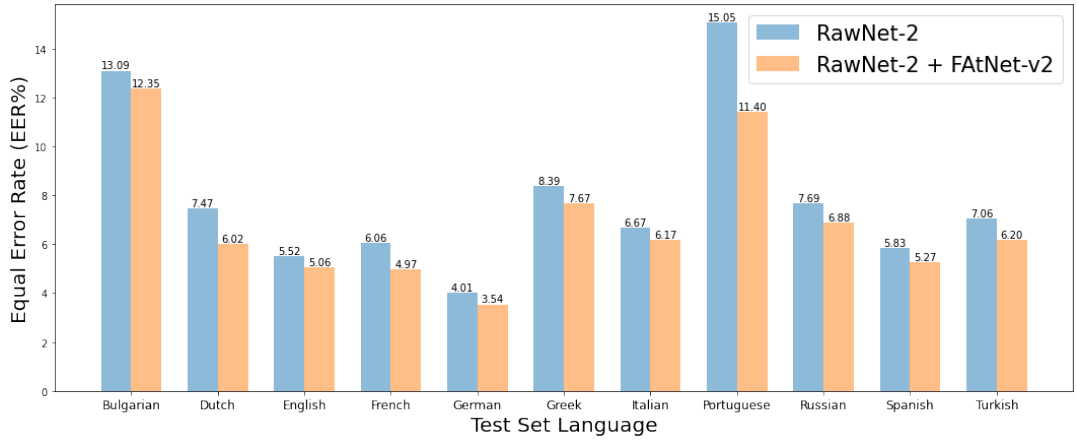



Figure 5.5: Figure illustrating that the integration of RawNet-2 with FAtNet-v2 consistently decreases the EER across test sets comprising speech in multiple languages.

### 5.5.3 Linguistic study with augmentation

To gain deeper linguistic insights, we analyze the performance of the standalone FAtNet models using test-time augmentation (TTA) for feeding input data. The input strategy outlined in Section 4.3 is referred to as $S_0$.

In the TTA strategy, each audio recording in the test set is either repeated or truncated to a fixed duration of 30 seconds [113], and then further divided into 3-second segments. All possible pairs of these segments are created, forming a batch of 100 pairs, which is fed to the model (similarly to $\mathbf{S_0}$). For FAtNet-v1, the final probabilities are averaged across the batch. For FAtNet-v2, the 100 embeddings of embedding$_1$ are

averaged to produce a single tensor, and the same is done for embedding$_2$. The cosine similarity between these averaged embeddings is then computed. This input procedure is referred to as $S_1$.

**Observations**: As shown in Figure 5.6, using the $S_1$ input strategy resulted in improved performance compared to $S_0$. Notably, FAtNet-v2 outperformed FAtNet-v1 on out-of-domain (Non-English) test sets. This is expected, as FAtNet-v2 employs two 4-head attention blocks, while FAtNet-v1 uses a single 8-head attention block. This suggests that selectively attending to frame-level features for each audio clip individually enhances the model's language robustness on out-of-domain data.

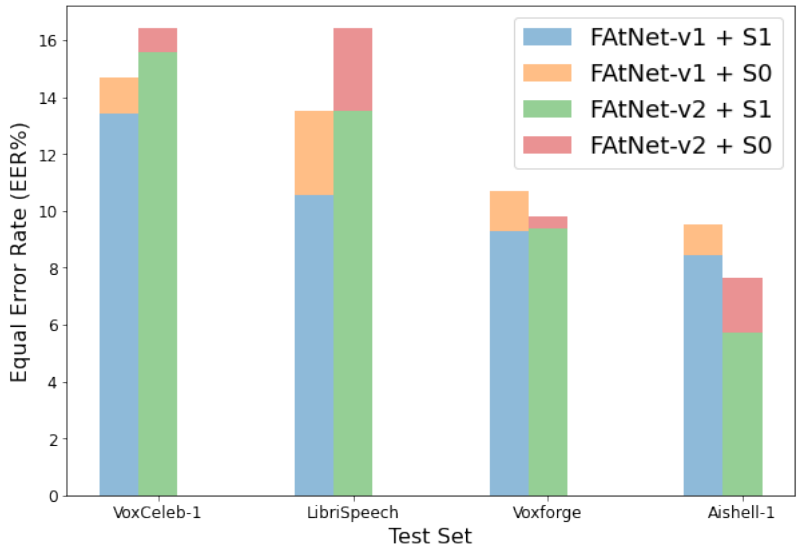



Figure 5.6: Figure illustrating the robustness of the proposed FAtNet models on out-of-domain test sets, demonstrating improved performance with the $S_1$ test-time augmentation strategy compared to $S_0$.

### 5.5.4 Qualitative comparison with the baselines

In this experiment, we assess the generalization performance of the proposed FAtNet networks compared to baseline models. As shown in Figure 5.7, the baselines performed reasonably well on the VoxCeleb-1 test set (mostly English) and the LibriSpeech (English) test set. However, their performance dropped noticeably on the two out-of-domain multilingual test sets: Aishell-1 (Mandarin) and Voxforge (Non-English).

In contrast, FAtNet models demonstrated improved performance on the out-of-domain multilingual test sets, generalizing effectively without any domain adaptation. This highlights the language dependency present in the VGG-M and RawNet-2 base-

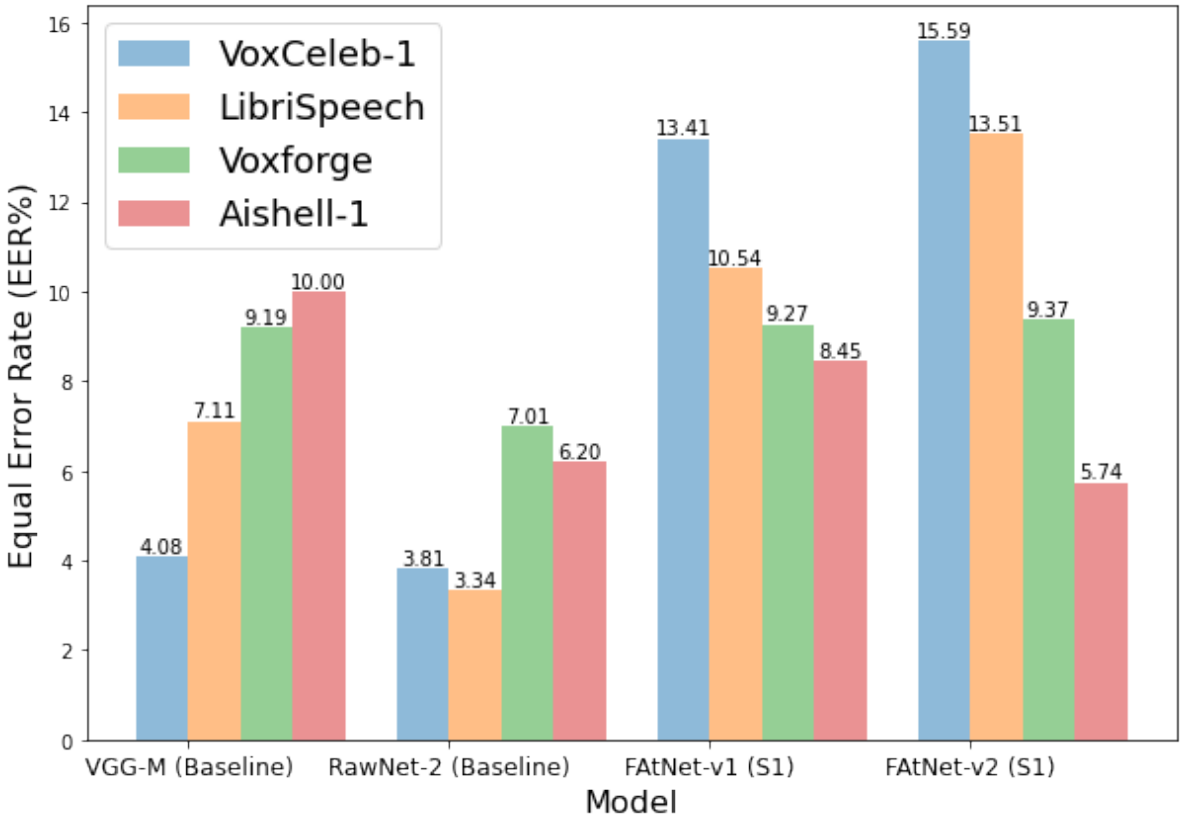



Figure 5.7: Figure illustrating that baseline models exhibit degraded performance on Non-English test sets without domain adaptation, whereas FAtNet models show improved performance on the same sets without any adaptation.

lines. The comparatively lower performance of FAtNet on VoxCeleb-1 and LibriSpeech may stem from dataset-specific variability: VoxCeleb-1 contains recordings collected under noisy, unconstrained conditions, while LibriSpeech, derived from audiobooks, exhibits high prosodic variation. These observations suggest that frame-attentive networks offer strong cross-lingual generalization but may be sensitive to noise or prosodic variability.

### 5.5.5 Ablation Study

To investigate which components of FAtNet contribute to its language robustness, we conducted an ablation study. FAtNet consists of two main modules: the TDNN component and the attention block. We trained a simple TDNN-only model on the VoxCeleb-2 dev set, keeping the architecture identical to FAtNet but omitting the attention block. Figure 5.8 and Figure 5.9 show that the full FAtNet models generally outperform the TDNN-only model across most test sets. These results indicate that frame-level features

enhance language robustness, and that attention-driven selection of these features further improves performance. Thus, combining the TDNN module with attention blocks, as in FAtNet, is key to achieving language-robust embeddings.

Interestingly, the TDNN-only model outperformed both FAtNet-v1 and FAtNet-v2 on the Aishell-1 and LibriSpeech test sets. While FAtNet-v1 surpassed TDNN on LibriSpeech, FAtNet-v2 did not, suggesting that FAtNet-v1 handles prosodic variations better. This is reasonable because FAtNet-v1 uses a single attention block applied after concatenating frame-level features from both audio clips, resulting in a higher-dimensional input to the attention mechanism than in FAtNet-v2. On the other hand, FAtNet-v2 performed better on out-of-domain test sets compared to TDNN, whereas FAtNet-v1 underperformed on the Aishell-1 test set relative to TDNN. The use of two separate attention blocks in FAtNet-v2, each specific to an individual audio clip, contributes to its superior language robustness compared to FAtNet-v1.

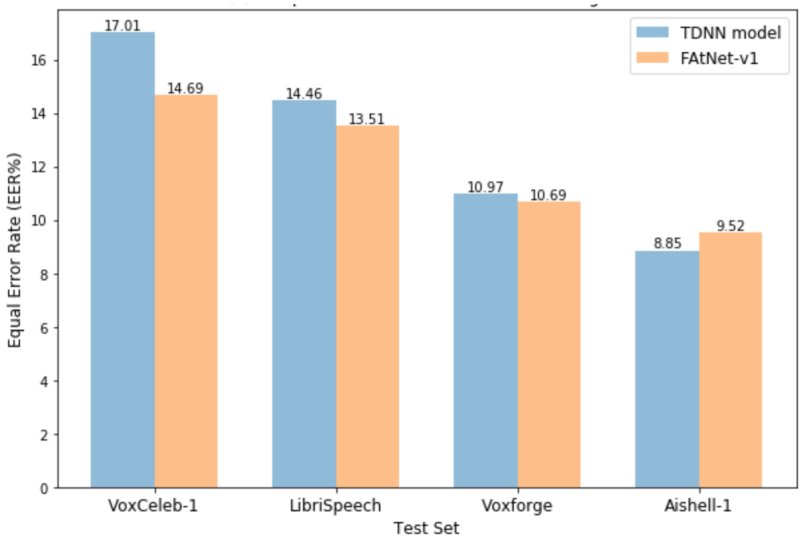



Figure 5.8: Comparison of the TDNN model and FAtNet-v1 performance, using the $S_0$ input strategy for both models.

## 5.6 Discussion

This study explores a cost-efficient approach of combining lightweight frame-attentive embeddings with more complex baseline models to reduce linguistic bias in these baselines without requiring domain adaptation. Through extensive experiments across twelve languages and ablation studies, we observed that the proposed method consistently and

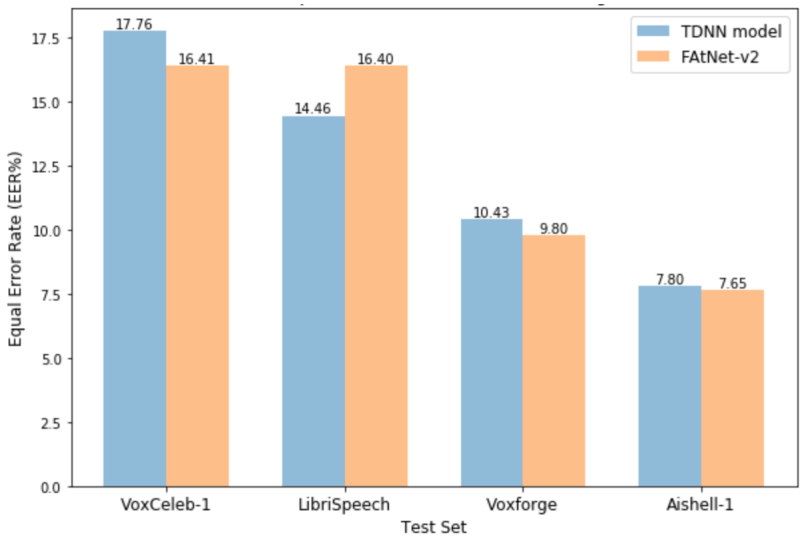



Figure 5.9: Comparison of the TDNN model and FAtNet-v2 performance, employing the $S_0$ input strategy for both models.

significantly improves the mitigation of linguistic bias. Key observations include:

**Model complexity:** As shown in Table 5.3, the FAtNet models contain fewer parameters compared to the baseline models, resulting in faster training times. While VGG-M and RawNet-2 require 71.7MB and 53.6MB of disk space, respectively, the proposed FAtNet-v1 and FAtNet-v2 occupy only 41MB and 32.6MB. Hence, FAtNet models are considerably lighter than the baselines.

| Model | #Parameters |
|---|---|
| VGG-M | 17909219 |
| RawNet-2 | 13379378 |
| **FAtNet-v1** | 10226690 |
| **FAtNet-v2** | 8127490 |

Table 5.3: Table presenting the parameter counts for the proposed FAtNet models and the baseline networks.

**Cost-effectiveness:** Reducing linguistic bias without domain adaptation is essential for improving the global usability of speaker verification models. However, this task is inherently challenging, often requiring complex network architectures and additional computational overhead. Some commonly used approaches to address linguistic bias include:

1. Fine-tuning a pre-trained model separately for each language can mitigate bias. Given that there are roughly 7,000 languages worldwide [59], this approach is extremely resource-intensive and impractical for large-scale deployment.

2. Training large, highly complex models on massive datasets can improve generalization to out-of-domain test sets. However, this strategy demands significant

computational resources and storage capacity.

In contrast, FAtNet models are lightweight and have lower complexity compared to these baselines. By integrating the proposed FAtNet embeddings with stronger and heavier baseline models, we observed substantial performance improvements. This indicates that with minimal overhead, FAtNet embeddings can effectively enhance the generalizability of existing baselines. Consequently, our approach provides a cost-efficient solution for mitigating linguistic bias and improving the usability of speaker verification models worldwide.

## 5.7 Conclusions and Future Work

In this chapter, we presented a cost-efficient approach that leverages lightweight frame-level embeddings to mitigate linguistic bias in existing speaker verification systems, avoiding the need for domain adaptation. We further examined the use of attention mechanisms at the frame level to selectively emphasize discriminative features. To rigorously evaluate our theoretical hypothesis, we proposed two variants of frame-attentive networks: FAtNet-v1 and FAtNet-v2, and studied their integration with baseline models across twelve languages. Experimental results demonstrated consistent improvements in baseline performance on out-of-domain test sets without any domain adaptation. Qualitative comparisons indicated that the proposed models are more generalizable than the baselines. Additionally, ablation studies revealed that frame-level embeddings capture less linguistic information than utterance-level embeddings, and intelligent selection of frame-level features can further enhance speaker verification performance.

Our findings also highlight directions for future research. For example, exploring standalone domain-invariant architectures could be promising. While this work focuses on trial pairs where both speech recordings are in the same language, extending the approach to bilingual or code-switched scenarios—where recordings in a pair come from different languages—may offer further benefits.

# CHAPTER 6

# Mitigating Partially Cross-Lingual Bias in Speaker Verification

Linguistic bias presents a major obstacle to ensuring diversity, equity, and inclusiveness in Natural Language Processing (NLP) systems. The problem is especially acute in security-critical applications such as speaker verification, where fairness and reliability are paramount. Speaker verification systems determine whether two speech recordings belong to the same speaker. Such biometric systems must be accurate and inclusive for bilingual users. Yet, deep neural network–based systems often exhibit linguistic bias. This bias can manifest as either fully cross-lingual or partially cross-lingual. While Chapter 5 addressed the issue of fully cross-lingual bias, this chapter focuses on the challenge of partially cross-lingual bias in speaker verification. Partially cross-lingual bias arises when one recording in a trial pair belongs to the training language and the other is in an unseen target language. Such mismatches can distort system decisions, creating barriers for bilingual speakers. Although domain adaptation can mitigate this issue, adapting models to each language individually is prohibitively expensive.

In this chapter[1], we explore cost-efficient strategies to reduce partially cross-lingual bias in speaker verification. We begin by analyzing the behavior of five baseline systems across five partially cross-lingual scenarios. Building on these insights, we propose *EcoSpeak*, a lightweight framework that incorporates contrastive linguistic (CL) attention. CL attention exploits linguistic differences within trial pairs to emphasize more discriminative regions of the embeddings. Experimental results demonstrate that EcoSpeak consistently improves robustness under partially cross-lingual testing, providing an efficient and practical solution toward fairer speaker verification.

[1]This chapter presents the following paper:
Divya V 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.

## 6.1 Introduction

Linguistic bias is a significant concern that undermines the diversity, equity, and inclusiveness of Natural Language Processing (NLP) systems. The problem becomes even more critical in security-sensitive applications such as speaker verification, where fairness and reliability are essential. Speaker verification is a biometric task that determines whether two speech recordings belong to the same speaker. These two recordings form a trial pair, labeled positive if they originate from the same speaker and negative otherwise. Speaker verification has broad applications in domains such as forensics, e-commerce, legal proceedings, and access-control systems [46]. Such systems can operate in either text-dependent or text-independent modes [180]. Text-independent speaker verification imposes no constraints on the spoken content, relying instead on acoustic features to discriminate between speakers. By eliminating the need for users to remember specific passphrases, text-independent approaches provide a more seamless and user-friendly experience compared to their text-dependent counterparts.

Deep Neural Network (DNN)–based methods have achieved state-of-the-art performance in text-independent speaker verification [31, 113, 114]. However, the embeddings extracted by these models often blend acoustic and linguistic information [202]. This entanglement introduces linguistic bias, causing models to rely on irrelevant language cues when verifying speakers, which in turn degrades performance on unseen target languages [95, 190]. Linguistic bias can manifest in two ways: fully cross-lingual or partially cross-lingual. In the fully cross-lingual case, both recordings in a trial pair belong to a target language $t$ that differs from the source or training language $s$. In the partially cross-lingual case, one recording is in $s$ while the other is in $t$, creating a mismatched condition that is equally challenging, as shown in Figure 6.1.

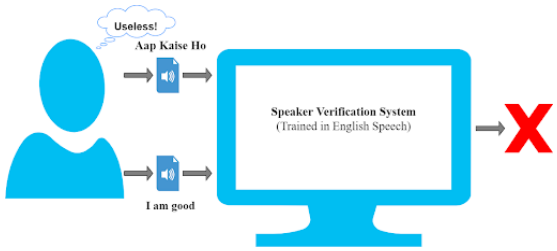



Figure 6.1: A partially cross-lingual scenario in speaker verification.

While most prior work has emphasized the fully cross-lingual setting, the partially cross-lingual scenario deserves equal attention, particularly since nearly 40% of the

world's population is bilingual [180]. Addressing this problem is essential to improve the inclusiveness of speaker verification technologies. Domain adaptation offers one solution [86, 203, 24, 177, 138, 166, 182], but adapting models to thousands of languages is impractical and prohibitively expensive. Another strategy is to train models on large-scale multilingual datasets to improve generalizability [28], but this requires immense computational and storage resources, which also translates into high carbon emissions and environmental costs [142, 187].

To address these challenges, this chapter investigates cost-efficient methods for mitigating partially cross-lingual bias in text-independent speaker verification. We introduce EcoSpeak, a lightweight framework designed to reduce this bias without expensive domain adaptation. EcoSpeak combines three key components: a compact residual network, a novel contrastive linguistic (CL) attention mechanism, and a bias corrector. Residual connections strengthen the model's ability to retain low-level acoustic cues that are critical for speaker verification. The CL attention mechanism leverages linguistic differences within trial pairs to generate attention weights, directing the model toward more discriminative features. Finally, the bias corrector adjusts verification probabilities based on the linguistic mismatch between recordings.

We first analyze the behaviour of five baseline systems across five partially cross-lingual test sets derived from four low-resource languages. We then evaluate EcoSpeak on these test sets, demonstrating its effectiveness even without domain adaptation. In addition, we explore efficient fine-tuning strategies that further improve EcoSpeak's generalizability to unseen low-resource languages, all while keeping computational costs low.

**Summary of Chapter Contributions:**

1. We study the behavior of five baseline speaker verification models on five partially cross-lingual test sets spanning four low-resource languages.
2. We introduce **EcoSpeak**, a cost-effective approach to mitigate bias in partially cross-lingual speaker verification.
3. We investigate the effectiveness of EcoSpeak on partially cross-lingual test sets and explore low-cost fine-tuning strategies to improve its generalizability to unseen languages.

## 6.2 Related Works

**Partially Cross-Lingual Bias:** Training on large-scale cross-lingual datasets can help reduce partially cross-lingual bias [180, 131]. However, such labeled datasets are often scarce [180], and using them entails significant computational and storage costs. Multi-task learning is another approach, allowing models to simultaneously learn speaker identities while alleviating linguistic bias [202]. Additionally, combining multiple models through fusion can further mitigate linguistic bias [131, 165], though this increases inference costs. Residual networks, in particular, have shown relatively higher robustness to linguistic differences compared to other architectures [131, 165], though the underlying reasons remain unclear. In this chapter, we examine the behavior of residual networks under partially cross-lingual conditions. Among prior studies, [165] is the most closely related. They address partially cross-lingual speaker verification involving speakers whose first language is Persian and second language is English [198], proposing a language compensation offset for trial pairs in different languages. However, their study focuses on closed-set speaker verification, where test utterances belong to known speakers. In contrast, we target the open-set scenario, where test trial pairs may include speakers unseen during training.

**Green Speech Processing:** The NLP community has increasingly emphasized developing inclusive and environmentally sustainable models [142, 187]. Speech processing, however, is resource-intensive, demanding substantial computation and storage. For example, SpeakerStew's training set included over 20 million utterances from 196,000 speakers [28], and XLS-R contains roughly 2 billion parameters, trained on nearly half a million hours of speech [10]. In [131], models were trained using recordings from 21,795 virtual speakers along with actual training set speakers for partially cross-lingual bias mitigation. Such large-scale training contributes to high carbon emissions. Consequently, researchers have explored cost-efficient bias mitigation techniques for fully cross-lingual speaker verification [89]. This chapter extends these efforts by investigating low-cost bias mitigation strategies for partially cross-lingual speaker verification.

## 6.3 Proposed Approach

The outstanding performance of deep Convolutional Neural Network (CNN)–based models in speaker recognition motivates an investigation into their susceptibility to linguistic bias. These models employ several CNN layers [113, 73, 74]. The initial layers typically capture low-level acoustic properties of speech, which are crucial for speaker verification [88], while the deeper layers capture higher-level representations that often embed linguistic information [116]. As a result, deeper models may inadvertently learn excessive linguistic details and become biased. To address this issue, we introduce EcoSpeak, a framework designed to mitigate linguistic bias in speaker verification. This section outlines the architecture of EcoSpeak.

**Hypothesis:** Residual connections in deep CNNs propagate outputs from lower layers to higher layers [58]. This mechanism integrates low-level acoustic features with more abstract higher-level representations, enabling the model to preserve and emphasize the acoustic cues critical for speaker verification. Based on this, we hypothesize that residual connections reduce linguistic bias by reinforcing the importance of low-level acoustic information.

**Input:** The trial pair recordings are first preprocessed by trimming silent regions. From the processed audio, we extract 64-dimensional normalized log-Mel spectrograms with shape $(b, 1, t_i, m)$, as illustrated in Figure 6.2. Here, $b$ refers to the batch size, $1$ corresponds to the mono-channel input, $t_i$ represents the number of time frames, and $m$ denotes the number of Mel bands ($m = 64$). Since test recordings can have variable durations, $t_1$ and $t_2$ may differ across inputs. These spectrogram features are then fed into ResNet (Lite) and the $s$-Detect model.

**$s$-Detect:** In partially cross-lingual trial pairs, one recording belongs to the source language $s$ while the other is from an unseen target language $t$. To identify whether a given recording is in $s$, we employ the $s$-Detect module. This model is composed of three bidirectional GRU layers (hidden size = 128) followed by a fully connected layer. As illustrated in Figure 6.2, it outputs both a probability score and a 256-dimensional language identification embedding ($lid_i$, $d = 256$).

**ResNet (Lite):** ResNet (Lite) is a compact, more efficient variant of ResNet-34

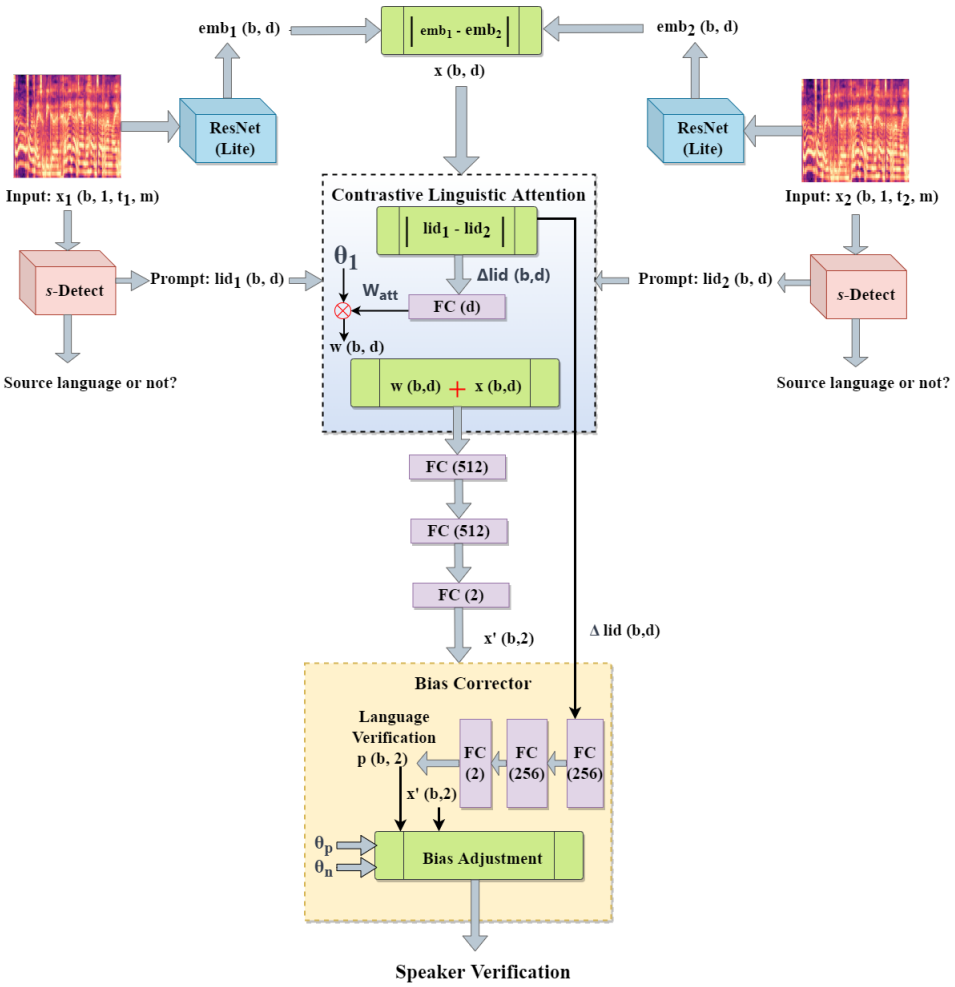



Figure 6.2: Architecture diagram for EcoSpeak.

[137].[2] Layer-wise details of ResNet (Lite) are provided in Table 6.1. The network is pre-trained for speaker identification, a multi-class task where the system predicts a speaker's identity from the training set. Within EcoSpeak, ResNet (Lite) is used to derive $d$-dimensional speaker embeddings (see Figure 6.2). Specifically, embeddings $emb_1$ and $emb_2$ for the trial pair are extracted from the avgPool layer. We then compute their absolute difference:

$$x = |emb_1 - emb_2| \tag{6.1}$$

This operation highlights the discriminative features crucial for speaker verification, while the use of absolute difference ensures commutativity, making the result invariant to the input order.

[2]Two variants of ResNet-34 were evaluated, and we selected the lighter yet robust one. Details are provided in the ablation study.

| Layer | Input shape | Output shape |
|---|---|---|
| conv1 | [b, 1, 301, 64] | [b, 32, 297, 60] |
| maxpool1 | [b, 32, 297, 60] | [b, 32, 149, 30] |
| layer1 | [b, 32, 149, 30] | [b, 32, 149, 30] |
| layer2 | [b, 32, 149, 30] | [b, 64, 38, 8] |
| layer3 | [b, 64, 38, 8] | [b, 128, 10, 2] |
| layer4 | [b, 128, 10, 2] | [b, 256, 3, 1] |
| avgpool | [b, 256, 3, 1] | [b, 256, 1, 1] |
| fc1 | [b, 256] | [b, 512] |
| fc2 | [b, 512] | [b, num_speakers] |

Table 6.1: Architecture details of the ResNet (Lite) speaker identification model.

**Contrastive Linguistic (CL) Attention:** Prior studies have shown that attention mechanisms are effective for speaker verification [38, 73]. Building on this, we introduce the contrastive linguistic (CL) attention mechanism specifically for partially cross-lingual speaker verification. The core idea is to leverage the linguistic mismatch between recordings in a trial pair to generate attention weights. The attention module takes $x$ as its primary input, while $lid_1$ and $lid_2$ serve as auxiliary prompts. The CL attention operates in two stages:

1. **Attention weight generation:** We begin by computing the absolute difference between the language embeddings of the two recordings. This difference is passed through a fully connected layer followed by a $ReLU$ activation to produce the CL attention weights $W_{\text{att}}$:

$$\Delta\text{lid} = |\text{lid}_1 - \text{lid}_2| \tag{6.2}$$

$$W_{\text{att}} = \text{ReLU}(\Delta\text{lid}\, W^T + b_{\text{linear}}) \tag{6.3}$$

   where $W$ and $b_{linear}$ are the weight and bias parameters of the linear layer.

2. **Attention application:** The generated attention weights are then applied to the speaker embedding difference. The final CL attention output is computed as:

$$x' = x + \tanh(\theta_1) \cdot W_{\text{att}} \tag{6.4}$$

   Here $\theta_1$ is a learned parameter.

**Bias Corrector:** To carry out speaker verification, $x'$ is first passed through fully connected layers (see Figure 6.2). The resulting speaker verification scores are then refined using the bias corrector, which operates in two stages: language verification and bias adjustment.

**Language verification:** EcoSpeak is jointly trained for both speaker and language

verification. Language verification is framed as a binary classification problem: determining whether the two recordings in a trial pair belong to the same language. For this task, $\Delta lid$ is passed through fully connected layers to produce the language verification probabilities $p$, as illustrated in Figure 6.2.

**Bias adjustment:** Speaker verification outcomes are often influenced by the linguistic similarity of the trial pair. When both recordings are in the same language, the system tends to lean toward the positive class, while cross-lingual pairs push the system toward the negative class. To counteract this tendency, EcoSpeak integrates a bias adjustment mechanism.

- If the language verification model predicts that both recordings are in the same language, the bias corrector penalizes the negative class to reduce the system's inclination toward the positive class:

$$x'[i,0] = x'[i,0] + |\theta_n| \tag{6.5}$$

- Conversely, if the recordings are predicted to be in different languages, the bias corrector penalizes the positive class to prevent the system's inclination towards the negative class:

$$x'[i,1] = x'[i,1] + |\theta_p| \tag{6.6}$$

Here, $\theta_p$ and $\theta_n$ are learnable parameters.

## 6.4 Experimental Setup

The datasets[3] and baseline models[4] employed in this study are publicly accessible. All sets maintain a balanced number of positive and negative trial pairs. In our experiments, English serves as the source language $s$, while Tamil, Telugu, Malayalam, and Kannada are treated as the low-resource target languages $t$.

[3]VoxCeleb:https://www.robots.ox.ac.uk/~vgg/data/voxceleb/, Indian-English (NPTEL): https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset, Hindi:http://openslr.org/103/, Tamil:http://openslr.org/65/, Telugu:http://openslr.org/66/, Malayalam:http://openslr.org/63/, Kannada:https://openslr.org/79/, Microsoft Speech Corpus:https://www.microsoft.com/en-za/download/details.aspx?id=105292, NISP:https://github.com/iiscleap/NISP-Dataset

[4]VGG-M: https://github.com/Derpimort/VGGVox-PyTorch, X-Vector:https://huggingface.co/speechbrain/spkrec-xvect-voxceleb, ECAPA-TDNN:https://huggingface.co/speechbrain/spkrec-ECAPA-voxceleb, RawNet-2:https://github.com/Jungjee/RawNet/tree/master/python/RawNet2, RawNet-3:https://github.com/Jungjee/RawNet/tree/master/python/RawNet3

### 6.4.1 Datasets

#### Pre-training ResNet (Lite)

ResNet (Lite) was pre-trained for speaker identification using the VoxCeleb-2 development set [113, 31], which comprises 1,092,009 utterances from 5,994 speakers. The model's performance was subsequently evaluated on the VoxCeleb-1 test set [114], containing 37,720 trial pairs. Since the VoxCeleb datasets primarily feature English speech [131], English serves as the source language ($s$) in our experiments.

#### Training $s$-Detect

The $s$-Detect model was trained on English and five Indian languages: Hindi, Tamil, Telugu, Malayalam, and Kannada. Indian-accented English recordings were sourced from the NPTEL 2020 lecture dataset [4], while Hindi speech came from the Multilingual and Code-Switching ASR Challenge Dataset (sub-task 1) [41]. For Tamil and Telugu, we used conversational speech recordings from the Microsoft Speech Corpus [105], and Malayalam and Kannada recordings were obtained from OpenSLR [57].

#### Training Setup

We first trained ResNet (Lite) for speaker identification using the VoxCeleb-2 dev set. Each epoch took approximately 40 minutes to complete, resulting in 0.18 kgCO2eq of carbon emissions and consuming 0.61 kWh of electricity. The model was trained for a total of ten epochs. Next, we trained the $s$-Detect model to identify the source language (English). We combined speech recordings from multiple datasets, obtaining 23,856 (Hindi), 24,884 (Tamil), 20,207 (Telugu), 1,983 (Malayalam), 3,633 (Kannada), and 74,563 (English) recordings.

Subsequently, $s$-Detect was fine-tuned on the NISP-Hindi speaker data to enable EcoSpeak-Hindi to adapt to dataset-specific variations. Mixed training was performed by combining NISP-Hindi data with the original $s$-Detect training set. The adapted $s$-Detect was then used to train EcoSpeak-Hindi on the NISP-Hindi speaker recordings. Fine-tuning was performed for four epochs to avoid overfitting due to limited data, with ResNet (Lite) weights frozen. CrossEntropyLoss was employed as the loss function,

and the Adam optimizer with a learning rate of 0.0005 was used.

The same procedure was applied to train EcoSpeak-Tamil, EcoSpeak-Telugu, EcoSpeak-Malayalam, and EcoSpeak-Kannada models. Training was conducted on a single NVIDIA A100 GPU. Pre-processing and feature extraction were performed using Librosa, model training used PyTorch, and carbon emissions as well as electricity usage were tracked with CodeCarbon [101, 126].[5]

**Cross-lingual Speaker Verification**

Cross-lingual experiments were conducted using the NISP dataset [76], which contains speech from bilingual speakers whose native language is Hindi, Tamil, Telugu, Malayalam, or Kannada, and who speak English as a second language. Each speaker in the dataset contributes recordings in both English and their native language.

### 6.4.2 Low-Resource Language Test Sets

Our experiments target four low-resource languages (LRLs): Tamil, Telugu, Malayalam, and Kannada. For cross-lingual evaluation, we used native speaker data from NISP-LRL. The following notations are used consistently throughout to report our results:

1. $s$: The source language, which is English.
2. $t$: The target language, i.e., the speaker's native language.
3. $ts$ or $st$: A trial pair in which one recording is in English ($s$) and the other is in the speaker's native language ($t$).

We constructed seven LRL test sets, represented using the following notations:

1. $tt - tt$: Both recordings in each trial pair are in the speaker's native language $t$.
2. $ts - tt$: Positive trial pairs consist of recordings in different languages $ts$, while negative trial pairs are in the speaker's native language $tt$.
3. $ts - ts$: Each trial pair includes recordings in different languages $ts$.
4. $tt - ts$: Positive trial pairs have both recordings in the speaker's native language $tt$, whereas negative trial pairs consist of recordings in different languages $ts$.
5. $ss - ss$: All recordings in the trial pairs are in English $s$.
6. $ss - st$: Positive trial pairs include recordings in English $ss$, and negative trial pairs have recordings in different languages $st$.

[5] https://pypi.org/project/codecarbon/

7. $st - ss$: Positive trial pairs contain recordings in different languages $st$, while negative trial pairs are in English $ss$.

Table 6.2 provides a concise overview of the seven LRL test sets. Each set comprises 100,000 trial pairs, with 25,000 pairs contributed by native speakers of each low-resource language. In negative trial pairs, the speakers are matched by gender. Based on this, we created a same-language test set ($ss - ss$), a fully cross-lingual test set ($tt - tt$), and five partially cross-lingual test sets ($ts - tt$, $ts - ts$, $tt - ts$, $ss - st$, and $st - ss$).

| Test Set | Positive Trial Pairs | Negative Trial Pairs |
|---|---|---|
| $\boldsymbol{tt - tt}$ | both target ($tt$) | both target ($tt$) |
| $\boldsymbol{ts - tt}$ | one target ($t$), one source ($s$) | both target ($tt$) |
| $\boldsymbol{ts - ts}$ | one target ($t$), one source ($s$) | one target ($t$), one source ($s$) |
| $\boldsymbol{tt - ts}$ | both target ($tt$) | one target ($t$), one source ($s$) |
| $\boldsymbol{ss - ss}$ | both source ($ss$) | both source ($ss$) |
| $\boldsymbol{ss - st}$ | both source ($ss$) | one source ($s$), one target ($t$) |
| $\boldsymbol{st - ss}$ | one source ($s$), one target ($t$) | both source ($ss$) |

Table 6.2: Overview of the Low-Resource Language (LRL) test sets. In this context, $s$ denotes the source language (English) and $t$ denotes the target language (the speaker's native language). The $tt$–$tt$ set corresponds to a fully cross-lingual scenario, while $ss$–$ss$ represents a same-language test set. The other five test sets reflect partially cross-lingual conditions.

### 6.4.3 Baselines

We analyzed the behavior of five baseline models on the LRL test sets: RawNet-3, ECAPA-TDNN, RawNet-2, X-Vectors, and VGG-M [74, 38, 134, 73, 152, 113]. These baselines were pre-trained for speaker identification and take speech recordings as input to produce speaker embeddings. For speaker verification, each trial pair recording is fed into the model, and the cosine similarity between the resulting embeddings is computed to determine whether the recordings belong to the same speaker. The X-Vector, ECAPA-TDNN, and RawNet-3 models were trained on the combined VoxCeleb-1 and VoxCeleb-2 development sets, while VGG-M and RawNet-2 were trained on VoxCeleb-1 dev and VoxCeleb-2 dev, respectively.

### 6.4.4 Evaluation Metric

Equal Error Rate (EER) is a widely used metric for evaluating speaker verification systems [53]. It corresponds to the point where the False Match Rate (FMR) equals the False Non-Match Rate (FNMR). FMR measures the proportion of negative trial pairs that the system mistakenly classifies as positive, whereas FNMR measures the proportion of positive trial pairs incorrectly classified as negative. The EER is identified at the threshold where FMR and FNMR are equal. In this study, EER is employed to assess system performance, with lower values indicating better verification accuracy.

## 6.5 Experiments and Results

### 6.5.1 Baseline Behavioral Insights

Understanding error patterns in baseline models is the first step toward mitigating bias [26]. To this end, we analyzed the behavior of baseline models on the LRL test sets. As shown in Table 6.3, elevated EER values were observed on the $ts - tt$ and $st - ss$ test sets, indicating that high linguistic similarity in negative trial pairs ($tt$ or $ss$) degrades performance. This suggests that high linguistic similarity biases the model toward predicting the positive class. Conversely, low linguistic similarity in positive trial pairs ($ts$ or $st$) also reduces performance, implying a bias toward the negative class. We further observed lower EER values on the $tt-ts$ and $ss-st$ test sets, indicating that baseline models perform best when positive trial pairs have high linguistic similarity ($ss$ or $tt$) and negative trial pairs have low linguistic similarity ($ts$ or $st$). Overall, these results highlight that linguistic similarity between trial pair recordings strongly influences baseline model decisions.

**Key Observations:**

1. Elevated EERs were found on $ts-tt$ and $st-ss$, indicating that baselines perform worst on these sets. This reflects that linguistic mismatch in positive trial pairs ($ts$ or $st$) and linguistic match in negative trial pairs ($tt$ or $ss$) leads to performance degradation. We classify Positive-$ts$, Positive-$st$, Negative-$tt$, and Negative-$ss$ as **complex trial pair** types.
2. Lower EERs were observed on $tt - ts$ and $ss - st$, indicating that baselines perform best on these sets. This shows that linguistic match in positive trial pairs

| Model | $tt-tt$ | $ts-tt$ | $ts-ts$ | $tt-ts$ | $ss-ss$ | $ss-st$ | $st-ss$ |
|---|---|---|---|---|---|---|---|
| **VGG-M** (Baseline) | 11.40 | 26.15 | 22.42 | 9.47 | 10.35 | **7.90** | **28.06** |
| **X-Vector** (Baseline) | 6.75 | 20.38 | 17.43 | 5.85 | 6.92 | **5.25** | **22.19** |
| **ECAPA-TDNN** (Baseline) | 12.46 | 20.93 | 19.57 | 11.96 | 11.40 | **9.30** | **22.65** |
| **RawNet-2** (Baseline) | 38.24 | **41.48** | 39.21 | **36.87** | 37.90 | 37.00 | 39.80 |
| **RawNet-3** (Baseline) | 41.34 | **52.17** | 46.54 | **36.75** | 41.71 | 44.10 | 43.60 |
| **ResNet+** (Hypothesis) | 10.72 | **13.55** | 12.27 | 9.81 | **9.51** | 9.55 | 12.08 |
| **EcoSpeak** (Scheme-A) | 8.54 | **13.88** | 12.80 | 7.64 | 7.70 | **7.37** | 13.66 |
| **EcoSpeak** (Scheme-B) | 7.70 | 12.01 | **12.65** | 8.09 | **7.23** | 7.61 | 11.87 |
| **EcoSpeak** (Scheme-C) | 7.31 | 9.32 | **11.16** | 9.06 | **6.81** | 8.18 | 9.65 |

Table 6.3: EER (%) of baselines, ResNet+, and EcoSpeak across the LRL test sets. Bold font highlights each model's best and worst performance. Key observations: 1) Baselines show the lowest performance on $ts$–$tt$ and $st$–$ss$ sets. 2) ResNet+ exhibits more stable performance than the baselines. 3) EcoSpeak (Scheme-C) performed the better than that of Scheme-A and Scheme-B models. EcoSpeak (Scheme-C) performed the worst on $ts$–$ts$, differing from the baselines' worst-case pattern.

($tt$ or $ss$) and linguistic mismatch in negative trial pairs ($ts$ or $st$) improves performance. We classify Positive-$tt$, Positive-$ss$, Negative-$ts$, and Negative-$st$ as **simple trial pair** types.

### 6.5.2 Behavior of Residual Connections

We next explored the effect of residual connections on cross-lingual evaluation by testing ResNet+ on the LRL sets. ResNet+ employs 64, 128, 256, and 512 channels in its first through fourth layers, whereas ResNet (Lite) uses 32, 64, 128, and 256 channels. To assess stability, we compared the relative differences between each model's highest and lowest EER scores on the LRL test sets. As shown in Table 6.3, ResNet+ exhibited an EER difference of 29.81% (i.e., $\left(\frac{13.55-9.51}{13.55} \times 100\right)$), which is substantially lower than that of VGG-M, X-Vector, and ECAPA-TDNN. The corresponding relative EER differences for VGG-M, X-Vector, and ECAPA-TDNN were 71.84%, 76.34%, and 58.94%, respectively, indicating that ResNet+ is more stable across LRL test sets.

We also compared the RawNet models' performance on VoxCeleb-1 versus the LRL test sets. RawNet-2 and RawNet-3 achieved EERs of 3.67% and 1.11% on VoxCeleb-1. However, their performance deteriorated considerably on NISP-LRL sets, with EERs increasing by over 30%. In contrast, ResNet+ achieved an EER of 9.97% on VoxCeleb-1, which is closer to its LRL test set results. Notably, although ECAPA-TDNN, RawNet-2, and RawNet-3 also incorporate residual connections, they still exhibited high linguistic bias on LRL sets, suggesting that residual connections alone are

insufficient for mitigating bias.

**Summary of Findings:** ResNet+ demonstrates reduced linguistic bias compared to baseline models, indicating that residual connections can help mitigate bias. Nevertheless, residual connections by themselves are not adequate for complete bias reduction.

### 6.5.3 Data Balancing Schemes

Focusing on the quality of training data rather than its sheer quantity can serve as a cost-efficient way to mitigate linguistic bias [159]. To study the role of data balancing in partially cross-lingual speaker verification, we experimented with three balancing strategies for fine-tuning EcoSpeak. These strategies differed in how they distributed simple versus complex trial pair types within the training set.

**Methodology:** We defined six trial pair categories for EcoSpeak fine-tuning: Positive-$ts$, Positive-$tt$, Positive-$ss$, Negative-$ts$, Negative-$tt$, and Negative-$ss$, where "positive" and "negative" indicate whether both recordings belong to the same speaker. The notations $tt$, $ss$, and $ts$ represent whether the trial pair is monolingual ($tt$, $ss$) or cross-lingual ($ts$). Based on our baseline analysis, Positive-$ts$, Negative-$tt$, and Negative-$ss$ are considered *complex* types, while Positive-$tt$, Positive-$ss$, and Negative-$ts$ are considered *simple* types. Using these, we designed three balancing schemes:

1. *Scheme-A:* 200,000 examples for each trial pair type.
2. *Scheme-B:* 250,000 examples for each complex type, and 150,000 for each simple type.
3. *Scheme-C:* 300,000 examples for each complex type, and 100,000 for each simple type.

Each scheme produced 1.2 million trial pairs, resulting in separate training sets. EcoSpeak was fine-tuned on NISP-Hindi speaker data using these scheme-specific sets, yielding three models—one per scheme. We then evaluated these models on the LRL test sets (Tamil, Telugu, Malayalam, Kannada) without domain adaptation.

**Observations:** We compared the spread between the highest and lowest EER values for each scheme. The relative differences were 46.90% ($\left(\frac{13.88-7.37}{13.88} \times 100\right)$), 42.84%, and 38.97% for Scheme-A, Scheme-B, and Scheme-C, respectively. Scheme-C thus produced the most consistent performance, likely because it emphasized complex trial

pair types in training. This highlights that effective data balancing can provide a cost-efficient pathway for bias mitigation. Notably, unlike baselines that performed worst on $ts$–$tt$ or $st$–$ss$, EcoSpeak (Scheme-C) showed its weakest performance on $ts$–$ts$, indicating a different error trend.

### 6.5.4 Dataset for fine-tuning EcoSpeak

Due to the scarcity of data in low-resource target languages, identifying suitable datasets for model fine-tuning is a major challenge. To address this, we investigated two strategies for fine-tuning EcoSpeak:

1. Fine-tuning with weakly related but diverse datasets
2. Fine-tuning with strongly related but limited datasets

**Methodology:** For this study, we selected Tamil as the target low-resource language ($t$). Using the LRL test sets, we constructed Tamil-LRL test sets by retaining only those trial pairs containing speech recordings of Tamil native speakers. This process yielded seven Tamil-LRL test sets, each containing 25,000 trial pairs.

For fine-tuning, we used two types of datasets: (i) NISP-Hindi, a relatively large and diverse dataset with 103 speakers, although Hindi is only weakly related to Tamil, and (ii) NISP-Telugu, NISP-Malayalam, and NISP-Kannada, each smaller (60 speakers) but linguistically closer to Tamil. Fine-tuning EcoSpeak on these datasets produced four variants: EcoSpeak-Hindi, EcoSpeak-Telugu, EcoSpeak-Malayalam, and EcoSpeak-Kannada.

**Observations:** As shown in Table 6.4, EcoSpeak-Hindi consistently achieved lower EER values on the Tamil-LRL test sets compared to the other variants. This indicates that fine-tuning on a weakly related but diverse dataset can be more effective than fine-tuning on a strongly related but limited dataset. The results also suggest that overfitting on small datasets restricts the model's ability to generalize.

### 6.5.5 Cost Analysis

This work investigates cost-efficient approaches to partially cross-lingual speaker verification by comparing the computational costs of baseline models with those of the

| Test Set | EcoSpeak-Hindi | EcoSpeak-Telugu | EcoSpeak-Malayalam | EcoSpeak-Kannada |
|---|---|---|---|---|
| $tt-tt$ | **8.31** | 9.70 | 9.98 | 10.36 |
| $ts-tt$ | **10.25** | 14.57 | 13.44 | 12.97 |
| $ts-ts$ | **11.42** | 15.94 | 15.78 | 14.34 |
| $tt-ts$ | **8.86** | 10.51 | 11.61 | 12.43 |
| $ss-ss$ | **6.26** | 8.18 | 9.05 | 9.85 |
| $ss-st$ | **7.42** | 10.26 | 11.04 | 12.46 |
| $st-ss$ | **8.94** | 12.63 | 12.49 | 11.17 |

Table 6.4: EER values (%) on Tamil-LRL test sets. The EcoSpeak model fine-tuned with NISP-Hindi data achieved the best performance. Although Hindi is only weakly related to Tamil, the NISP-Hindi dataset is more diverse.

proposed EcoSpeak. The comparison considers model size, parameter count, and inference costs (time, carbon emissions, and electricity usage). As shown in Table 6.5, EcoSpeak requires fewer parameters and has a smaller model size than the baselines. We further compared the inference costs of EcoSpeak-Hindi with those of the baselines on the $tt-tt$ LRL test set. Results indicate that EcoSpeak-Hindi achieves faster inference than most baselines, while also producing lower carbon emissions and consuming less electricity. Although EcoSpeak's inference cost is close to that of the X-Vector model, Table 6.3 shows that EcoSpeak is notably more stable. Specifically, EcoSpeak (Scheme-C) exhibits a relative EER variation of 38.97%, whereas X-Vector shows a much larger variation of 76.34%. These findings suggest that EcoSpeak offers a more cost-efficient and stable solution for partially cross-lingual speaker verification.

| Model | #Parameters | Size (MB) | Time (sec) | $CO_2$ (kgCO2eq) | Electricity (kWh) |
|---|---|---|---|---|---|
| **RawNet-3** | 16,280,322 | 62.30 | 4000 | 0.46 | 0.73 |
| **ECAPA-TDNN** | 22,150,912 | 85.00 | 2195 | 0.23 | 0.36 |
| **RawNet-2** | 13,379,378 | 51.10 | 1360 | 0.13 | 0.20 |
| **VGG-M** | 17,909,219 | 68.40 | 1252 | 0.11 | 0.18 |
| **X-Vector** | 8,172,473 | 31.50 | **1014** | **0.09** | **0.14** |
| **EcoSpeak** | **6,660,233** | **25.50** | 1165 | 0.10 | 0.16 |

Table 6.5: Table comparing the computational costs of EcoSpeak and baseline models. The reported model size and parameter count for EcoSpeak include those of $s$-Detect. The time, carbon emissions, and electricity usage reflect inference costs measured on the $tt-tt$ LRL test set.

### 6.5.6 Absolute Difference in EcoSpeak

EcoSpeak employs the absolute difference operation to compare trial pair embeddings $(emb_1, emb_2)$ for speaker verification. This section details the experiment that motivated this choice. We evaluated the following models:

**ResNet (Lite)-Concat (RC):** In this model, each trial pair recording is passed through ResNet (Lite) to obtain 256-dimensional embeddings $(emb_1, emb_2)$. The embeddings are concatenated to form a 512-dimensional vector, which is then fed through two fully connected layers of 512 units each. A final fully connected layer with two units outputs the speaker verification decision. The RC model requires 22.4 MB of storage.

**ResNet (Lite)-AbsoluteDifference (RAD):** Here, each trial pair recording is also processed through ResNet (Lite) to generate 256-dimensional embeddings $(emb_1, emb_2)$. The absolute difference between these embeddings is computed to produce a 256-dimensional vector, which is passed through two fully connected layers of 256 units, followed by a two-unit output layer for speaker verification. The RAD model occupies 20.9 MB of storage.

**Language-Specific LRL Test Sets:** For evaluation, we created separate test sets for each target low-resource language, resulting in Tamil-LRL, Telugu-LRL, Malayalam-LRL, and Kannada-LRL sets. These are subsets of the original LRL test sets described in Section 6.4.1 and contain trial pairs exclusively from native speakers, with 25,000 pairs per language.

**Observations:** Performance comparisons between RC and RAD on both the LRL test sets (Section 6.4.2) and language-specific LRL sets (Table 6.6) indicate that RAD consistently achieves lower EER values. This result motivated the adoption of the absolute difference operation in EcoSpeak.

### 6.5.7 Ablation Study

To evaluate EcoSpeak, we conducted an ablation study summarized in Table 6.7. First, we observed that *ResNet (Lite)* outperforms ResNet+ on the LRL test sets while being computationally lighter. Consequently, ResNet (Lite) was used in EcoSpeak to extract speaker embeddings from trial pair recordings. In *ResNet (Lite)+fc*, instead of cosine similarity, fully connected layers were used for speaker verification. The absolute differences between trial pair embeddings from ResNet (Lite) were fed to the fully connected layers, which were fine-tuned on NISP-Hindi native speaker data using Scheme-C. The subpar performance of ResNet (Lite)+fc indicates that data balancing

| Dataset | Test Set | RC | RAD |
|---|---|---|---|
| NISP-LRL | $tt-tt$ | 10.11 | **8.16** |
| | $ts-tt$ | 13.26 | **12.91** |
| | $ts-ts$ | **11.96** | 12.08 |
| | $tt-ts$ | 9.15 | **7.78** |
| | $ss-ss$ | 8.70 | **7.88** |
| | $ss-st$ | 9.96 | **8.32** |
| | $st-ss$ | 11.30 | 11.30 |
| NISP-Tamil | $tt-tt$ | 12.38 | **10.72** |
| | $ts-tt$ | **14.98** | 15.29 |
| | $ts-ts$ | **13.47** | 14.69 |
| | $tt-ts$ | 11.14 | **10.24** |
| | $ss-ss$ | 8.94 | **8.11** |
| | $ss-st$ | 9.69 | **8.88** |
| | $st-ss$ | **12.39** | 12.66 |
| NISP-Telugu | $tt-tt$ | 9.61 | **7.41** |
| | $ts-tt$ | 10.81 | **9.83** |
| | $ts-ts$ | 10.07 | **9.58** |
| | $tt-ts$ | 8.98 | **7.15** |
| | $ss-ss$ | 10.10 | **9.22** |
| | $ss-st$ | 10.62 | **8.04** |
| | $st-ss$ | 8.71 | **8.46** |
| NISP-Malayalam | $tt-tt$ | 7.24 | **6.36** |
| | $ts-tt$ | 11.21 | **11.12** |
| | $ts-ts$ | **9.28** | 9.69 |
| | $tt-ts$ | 6.17 | **5.61** |
| | $ss-ss$ | **6.56** | 6.79 |
| | $ss-st$ | 7.68 | **7.58** |
| | $st-ss$ | **8.99** | 9.44 |
| NISP-Kannada | $tt-tt$ | 10.64 | **7.97** |
| | $ts-tt$ | 15.61 | **15.26** |
| | $ts-ts$ | 14.05 | **13.97** |
| | $tt-ts$ | 9.78 | **7.98** |
| | $ss-ss$ | 8.12 | **7.32** |
| | $ss-st$ | 10.55 | **8.50** |
| | $st-ss$ | **14.13** | 14.39 |

Table 6.6: EER (%) values for RC and RAD across various LRL test sets. RAD achieves lower EERs than RC on most test sets, supporting the use of the absolute difference operation in EcoSpeak.

alone is insufficient for mitigating linguistic bias. Following this, we fine-tuned the CL attention model and EcoSpeak using Scheme-C, as described in Section 6.5.3.

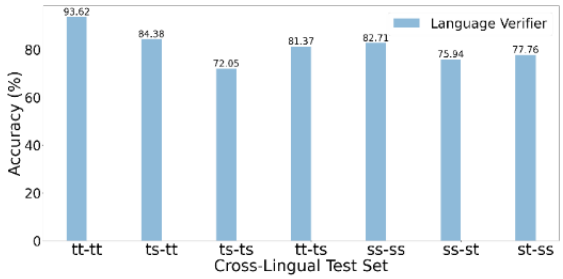
It is worth noting that the relative EER difference with ResNet+ is 29.81% as discussed in Section 6.5.2. This value is lower than that with the *CL attention*. However, the CL attention model surpassed ResNet (Lite) across most LRL test sets, with improvements of more than 20% through the minor cost of CL attention. Notably, using the CL attention model, we observe improvements in the two most challenging partially cross-lingual scenarios: $ts - tt$ and $st - ss$. This suggests that CL attention effectively modulates speaker embeddings based on the linguistic differences in trial pairs. *EcoSpeak*, which integrates both CL attention and the bias corrector, outperformed the CL attention model on most LRL test sets. Its lowest performance occurred on $ts - ts$, which can be explained by the interdependence of language verification and speaker verification. As shown in Figure 6.3, higher language verification accuracy corresponds to lower EER in speaker verification and vice versa. Since EcoSpeak performed worst on language verification for $ts - ts$, this accounts for its lowest speaker verification performance on that set.

| Model | $tt-tt$ | $ts-tt$ | $ts-ts$ | $tt-ts$ | $ss-ss$ | $ss-st$ | $st-ss$ |
|---|---|---|---|---|---|---|---|
| **ResNet+** | 10.72 | **13.55** | 12.27 | 9.81 | **9.51** | 9.55 | 12.08 |
| **ResNet (Lite)** | 9.52 | **12.14** | 10.96 | **8.33** | 8.54 | 9.13 | 10.58 |
| **ResNet (Lite)+fc** | 11.16 | **14.72** | 13.87 | 10.67 | 10.57 | **10.29** | 13.76 |
| **CL Attention** | 7.47 | 9.29 | **11.70** | 9.67 | **6.94** | 8.48 | 9.88 |
| **EcoSpeak** | 7.31 | 9.32 | **11.16** | 9.06 | **6.81** | 8.18 | 9.65 |

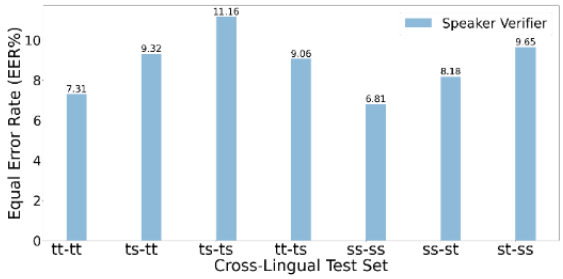
Table 6.7: Ablation study results for EcoSpeak. Observation: CL attention mitigates linguistic bias.

## 6.6 Conclusions and Future Work

This study examines the behavior of five baseline speaker verification models on five partially cross-lingual test sets. The results reveal that high linguistic similarity in negative trial pairs and low linguistic similarity in positive trial pairs lead to performance degradation. Additionally, residual networks show relative robustness under cross-lingual evaluation. Building on these insights, we introduced EcoSpeak, a cost-efficient approach to mitigating bias in partially cross-lingual speaker verification. EcoSpeak integrates residual connections, contrastive linguistic attention, and a bias corrector. Empirical evaluations demonstrate its robustness across partially cross-lingual test sets,




a) Language Verification




b) Speaker Verification

Figure 6.3: Negative correlation between EcoSpeak's language and speaker verification performance. Higher accuracy in language verification corresponds to lower EER in speaker verification, and vice versa.

with performance patterns differing from the baselines. The findings suggest that leveraging linguistic differences to selectively emphasize or suppress parts of speaker embeddings can effectively reduce cross-lingual bias.

These insights can guide the design of more robust, domain-invariant architectures. Moreover, this work encourages the adoption of greener methods in computationally intensive speech processing. For instance, our results indicate that fine-tuning on diverse datasets in a weakly related language can be more effective for bias mitigation in unseen low-resource target languages. The proposed data balancing strategies also reduce training costs compared to large-scale datasets. A thorough cost analysis is recommended to develop environmentally friendly and inclusive models.

## 6.7 Limitations

While this work proposes cost-efficient techniques for mitigating bias in partially cross-lingual speaker verification, the approach has the following limitations:

1. **Dependency on language verification performance:** EcoSpeak's speaker verification performance is influenced by its language verification accuracy. A more robust $s$-Detect model could enhance EcoSpeak's speaker verification results,

since $s$-Detect embeddings serve as prompts for CL attention and language verification.

2. **Need for broader validation of CL Attention:** The contrastive linguistic (CL) attention mechanism is based on the premise that the attention weights should align with speaker verification embeddings, emphasizing embedding components affected by linguistic variations. Although we evaluated CL attention on five partially cross-lingual test sets across four low-resource languages, additional experiments across more languages are necessary to fully validate its effectiveness.

3. **Limited target language data for $s$-Detect training:** EcoSpeak was not explicitly fine-tuned on the target low-resource languages (Tamil, Telugu, Malayalam, Kannada). Instead, $s$-Detect was trained using speech data from diverse datasets in these languages. This approach is practical, as language identification datasets are more readily available than cross-lingual bilingual speaker datasets for speaker verification.

Linguistic bias is a complex challenge to address using a single bias mitigation technique. EcoSpeak combines multiple low-cost strategies for bias reduction. Future work could explore integrating these techniques with additional methods advancing towards developing more inclusive and domain-invariant architectures.

# CHAPTER 7

# Quality and Sustainability Metrics for Large-Scale Audio Deepfake Detection and Anti-Spoofing Dataset Creation

Large-scale multilingual synthetic speech datasets are essential for advancing research in audio deepfake detection (ADD) and anti-spoofing, particularly for mitigating linguistic biases. To support this direction, Chapter 2 introduced the 4000-hour IndicSynth dataset. Subsequently, Chapter 3 demonstrated the vulnerability of human listeners to native-language-based audio deepfake attacks. Beyond linguistic bias, ADD and anti-spoofing systems are also affected by biases associated with accents and synthetic speech generation models. Consequently, there is a pressing need for large-scale synthetic speech datasets that encompass diverse languages, accents, speech generation models, and other demographic attributes, such as age, to support bias mitigation research.

The creation of such datasets is hindered by three key challenges. First, creating these resources requires rigorous quality assessments. However, current assessments primarily rely on human evaluation, which lacks scalability. Second, synthetic speech dataset creation often requires fine-tuning generation models for target languages and accents. The financial cost and resources incurred in fine-tuning undermine research inclusivity by deterring researchers with limited computational resources from participating in resource creation initiatives. Third, large-scale synthetic speech generation can incur substantial carbon emissions. While the environmental impact of speech processing has received limited attention in the literature, Chapter 6 of this thesis discussed the concept of green speech processing in the context of speaker verification. Similarly, modern generation models contain millions of parameters and incur significant computational cost. Therefore, sustainability considerations are crucial when generating synthetic speech on a large scale.

Given these challenges and the large number of publicly available synthetic speech generation models, selecting an appropriate model for dataset creation is non-trivial. In

particular, identifying models that are cost-efficient, environmentally sustainable, and suitable for large-scale dataset creation without additional fine-tuning remains challenging. To address this research gap, this chapter introduces *GreenVoice*, an automated framework for large-scale evaluation of synthetic speech quality. GreenVoice incorporates five evaluation metrics: Realism, Similarity, Environmental Impact Assessment (EIA), Cloning Quality Assessment (CQA), and G-Score. These metrics are designed to facilitate the selection of cost-efficient and environmentally sustainable speech generation models for large-scale synthetic speech dataset creation covering multiple demographic attributes. We demonstrate the effectiveness of GreenVoice through a case study covering eight synthetic speech generation models and seven English accents.

## 7.1 Introduction

Advances in synthetic speech generation technologies have made it possible to generate highly realistic voice clones from just a few seconds of reference audio [162]. This capability increases the risk of audio spoofing attacks, particularly for public figures whose voices are publicly available [79, 129]. Over the past three years, synthetic speech attacks have grown by more than 2000% [118], highlighting the urgent need for robust audio deepfake detection (ADD) and anti-spoofing systems [15]. Deploying these safeguards in smartphones, audio-based large language models, and other vulnerable platforms can help mitigate the risks posed by such attacks [15, 91].

Developing effective ADD and anti-spoofing models requires large-scale synthetic speech datasets [130]. Most publicly available datasets, however, are concentrated on high-resource languages such as English and Chinese [79, 47, 9, 108, 112], leaving many low-resource languages largely underrepresented [9, 194]. As a result, existing ADD and anti-spoofing models are often highly tailored to their training data [111], and even small linguistic variations in the test set can increase attack success rates by over 60% [118]. Systematic evaluations are therefore mostly limited to high-resource languages due to the scarcity of data in other languages [9], leaving cross-lingual biases and native-language-based deepfake vulnerabilities underexplored [9, 129].

Besides linguistic bias, ADD and anti-spoofing models also show generation-model-specific bias [118]. As a result, ADD and anti-spoofing models trained on data from one

generation model often perform poorly on samples from unseen models [78, 84, 109, 196]. With new synthetic speech generation models continually emerging, it is therefore essential to regularly expand ADD and anti-spoofing datasets to include a wider range of languages, accents, and generation methods [94, 130, 32, 157].

The development of large-scale ADD and anti-spoofing datasets is hindered by three key challenges, depicted in Figure 7.1. First, assessing data quality largely depends on human evaluations, which are both costly and difficult to scale. These evaluations also require participants fluent in the target languages and accents, restricting dataset design to languages and accents for which such participants are available. Consequently, many languages and accents remain underrepresented in current resources.

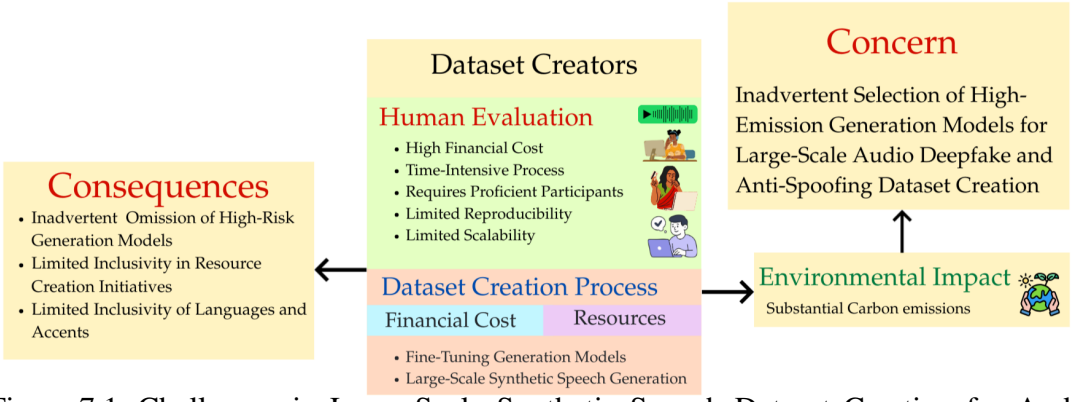



Figure 7.1: Challenges in Large-Scale Synthetic Speech Dataset Creation for Audio Deepfake Detection and Anti-Spoofing Research

The second challenge is that creating large-scale synthetic speech datasets with state-of-the-art generation models demands significant computational resources, including GPUs. Resource limitations often make it difficult for researchers to run these large models [70], and the challenge is even greater when datasets require training or fine-tuning models for specific languages [142, 79, 9]. As a result, contributing to large-scale dataset creation is extremely challenging for many researchers, limiting inclusivity and preventing those with constrained computational resources from participating in dataset development [129].

The third challenge concerns the substantial carbon emissions associated with large-scale synthetic speech generation. Although environmental assessments are rarely discussed in the speech processing literature, studies indicate that computational demands in deep learning have increased by 300,000 times over the past six years [142]. Training a large NLP model is estimated to generate approximately 626k lbs of $CO_2$ [139]. This amount of emissions is roughly five times the lifetime emissions of an average passen-

ger vehicle in the United States. Consequently, reducing carbon emissions by half over the next decade is necessary to mitigate the rising frequency of natural disasters [164]. Therefore, incorporating sustainability considerations into the creation of large-scale synthetic speech datasets is essential.

To address these challenges, reliable automated metrics are needed to enable comprehensive and large-scale comparative evaluations of synthetic speech generation models. Without such evaluations, models capable of producing high-quality synthetic speech may be unintentionally excluded during dataset creation. These models could subsequently be exploited by malicious actors. Additionally, large-scale dataset creation may incur substantial financial, computational, and environmental costs if costly generation models that offer only marginal improvements in output quality are selected.

This chapter introduces *GreenVoice*, an automated framework designed for large-scale comparative evaluation of synthetic speech generation models. The framework employs five objective metrics to assess synthetic speech quality automatically: *Realism Score*, *Similarity Score*, *Environmental Impact Assessment (EIA) Score*, *Cloning Quality Assessment (CQA) Score*, and *G-Score*. The EIA score measures the carbon emissions associated with synthetic speech generation. CQA and G-Score are composite metrics: CQA is a performance-oriented metric that combines the Realism and Similarity Scores into a single measure, whereas G-Score is an efficiency-oriented metric that integrates the CQA and EIA scores of generation models. We demonstrate the efficacy of GreenVoice through a comprehensive case study involving eight synthetic speech generation models and seven English accents.

**Summary of Chapter Contributions:**

1. We propose GreenVoice, an automated comparative evaluation framework for synthetic speech generation models. GreenVoice evaluates each model by jointly assessing its synthetic speech quality and environmental impact, producing a GreenVoice Score. Higher scores indicate models that are both efficient and environmentally sustainable.
2. We apply GreenVoice to evaluate five text-to-speech (TTS) and three voice conversion (VC) models across seven English accents, considering multiple evaluation criteria: (a) realism of synthetic voices, (b) similarity between cloned and bonafide target voices, (c) model size (parameter count), (d) inference or cloning time, (e) carbon emissions, and (f) electricity consumption.
3. Using the GreenVoice Score, we evaluate the eight models for two practical scenarios: (A) prioritizing high realism (authenticity) of synthetic voices, and (B)

prioritizing high similarity (mimicry) between cloned and bonafide voices.

4. We further analyze the models' performance on both in-domain and out-of-domain (OOD) accents, investigate potential gender biases, and provide recommendations for selecting efficient and sustainable generation models for audio deepfake detection (ADD) and anti-spoofing dataset creation.

## 7.2 Related Works

### 7.2.1 Challenges in Human Evaluation

At present, human evaluation is widely regarded as the gold standard for assessing synthetic speech quality [140, 136]. However, these evaluations are costly and difficult to scale across numerous languages and generation models [87, 35, 97]. In addition, human evaluations often lack reproducibility [29]. In response to these challenges, existing ADD and anti-spoofing datasets typically employ recently proposed generation models that support the target languages [79]. These datasets serve as valuable benchmarks for research. However, systematic comparative evaluation of generation models remains essential to identify high-risk models.

### 7.2.2 Reliability Challenges in Existing Automated Metrics

Mean Opinion Score (MOS) predictors were introduced to address the scalability limitations of human evaluations [87, 60]. These predictors enable large-scale automated evaluation of synthetic speech quality. However, they often show limited generalizability when applied to out-of-domain datasets [35]. Adapting MOS predictors to new domains requires extensive listening tests involving human participants [97]. Furthermore, Word error rate (WER) has also been used to evaluate synthetic speech quality [6]. However, calculating WER requires automatic speech recognition (ASR) systems, which often lack sufficient support for low-resource languages and accents [23]. As a result, WER cannot reliably assess synthetic speech quality in underrepresented languages [169]. This limitation highlights the need for language-agnostic metrics that enable large-scale automated evaluation of synthetic speech quality across diverse languages and accents, thereby improving inclusivity in evaluation practices [189].

DNSMOS is one such automated metric used for speech quality assessment [136]. However, it was originally designed to rank Deep Noise Suppression (DNS) techniques. In addition, prior work has examined the proximity between the Speech Activity Ratio (SAR) and the Signal-to-Noise Ratio (SNR) of bonafide and synthetic samples for quality evaluation [32]. These metrics measure distributional similarity between bonafide and synthetic audio in terms of silence patterns (SAR) and noise characteristics (SNR), rather than directly capturing perceptual realism.

### 7.2.3 Need for Joint Evaluation of Realism and Similarity

Beyond these metrics, ClonEval proposes an evaluation protocol for assessing synthetic speech quality [29]. This protocol measures the proximity between WavLM embeddings of bonafide and cloned (synthetic) speech. Although effective for evaluating similarity, it does not assess the naturalness of synthetic speech. Furthermore, there is an increasing interest in jointly optimizing ADD and anti-spoofing models. Therefore, there is a need for a reliable evaluation metric that jointly considers the realism and similarity dimensions.

### 7.2.4 Sustainability Considerations in Synthetic Speech Generation

Besides quality, the environmental impact of large-scale synthetic speech generation must also be considered [70]. However, most existing evaluation metrics focus only on quality and ignore the associated costs. Although some studies examine the computational cost of these models, their environmental impact remains largely underexplored. Synthetic speech generation models often require substantial computational resources, raising concerns about their environmental footprint [124]. Therefore, incorporating planetary boundary considerations into evaluation frameworks is important to ensure the sustainability of large-scale synthetic speech generation [164]. To address this, prior work has used Pareto-optimal analysis to study the trade-off between synthetic speech quality and environmental impact [42, 124]. However, Pareto-optimal approaches become difficult to scale when a large number of generation models must be evaluated.

### 7.2.5 Research Gap Addressed

Given the limitations of existing evaluation metrics, there is a clear need for a unified metric that jointly evaluates both the performance and environmental impact of synthetic speech generation models. To date, Pareto-optimal analysis is the closest approach explored in this direction, yet a unified evaluation metric remains lacking. To address this gap, this work introduces GreenVoice, an automated comparative evaluation framework for synthetic speech generation models. The framework incorporates five automated metrics to assess synthetic speech quality across realism, similarity, and sustainability dimensions. In particular, the G-Score jointly measures model performance and environmental sustainability, while the Cloning Quality Assessment (CQA) score is a performance-oriented metric that evaluates realism and similarity together.

## 7.3 GreenVoice Framework

GreenVoice integrates both cloning quality and environmental impact into a unified evaluation metric, the GreenVoice Score ($G\text{-}Score$). This section presents a detailed overview of the framework, as illustrated in Figure 7.2.

GreenVoice integrates cloning quality and environmental impact to compute a novel composite evaluation metric, the GreenVoice Score ($G\text{-}Score$). This section presents a detailed overview of the framework.

**Target Models:** GreenVoice is designed to evaluate text-to-speech (TTS) and voice conversion (VC) models, collectively referred to as target models. For each of the $T$ target models, GreenVoice computes a $G\text{-}Score$, which combines a Cloning Quality Assessment (CQA) score and an environmental impact assessment score. The CQA score integrates results from both realism and similarity tests.

**Realism Test:** The realism test evaluates how human-like the synthetic audio generated by the target models is. This test leverages state-of-the-art audio deepfake detection (ADD) models, which classify input recordings as either bonafide (real) or synthetic (cloned). For each target model $t_i$, an ADD test set is prepared containing $d$ audio clips with equal numbers of real and cloned samples. The test set is processed by an ADD model to compute an Equal Error Rate (EER) in ADD ($\mathrm{EER}^{i}_{\mathrm{ADD}}$) for $t_i$. A higher EER

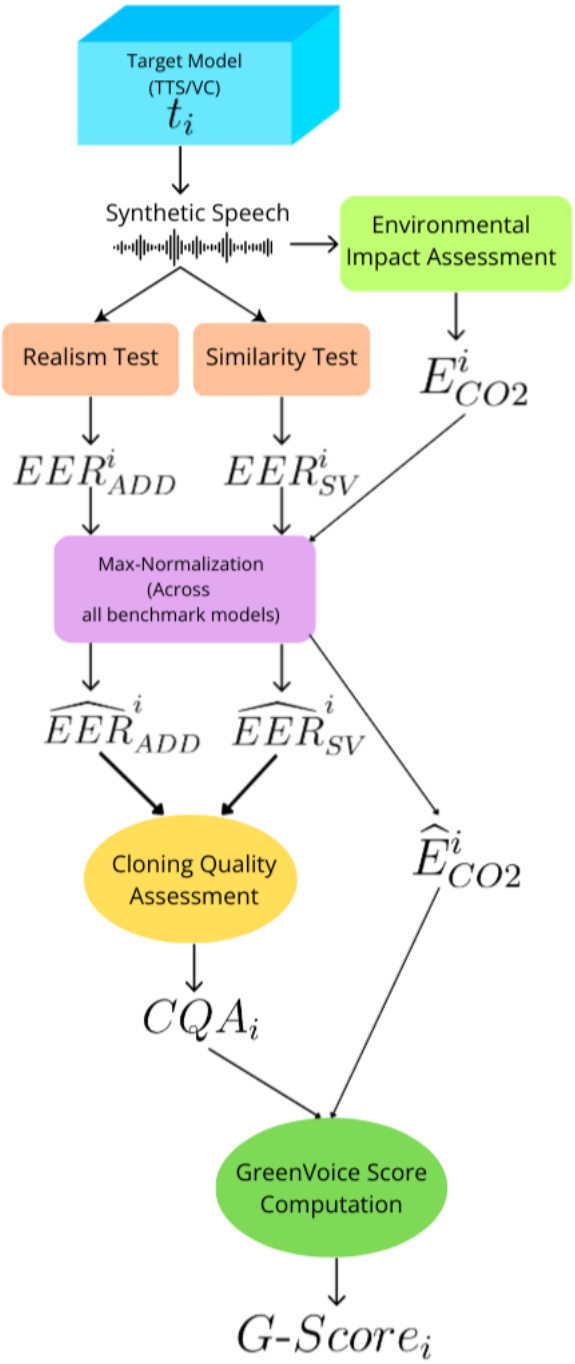



Figure 7.2: The GreenVoice framework evaluates target models by calculating each model's GreenVoice Score ($G\text{-}Score$), where a higher score reflects better overall performance. For a given model $t_i$, $G\text{-}Score_i$ integrates the Cloning Quality Assessment ($CQA_i$) and the Environmental Impact Assessment ($\widehat{\mathrm{E}}^i_{\mathrm{CO2}}$). $CQA_i$ combines results from realism and similarity tests for $t_i$, while $\widehat{\mathrm{E}}^i_{\mathrm{CO2}}$ measures the carbon emissions generated during voice cloning. By uniting performance and sustainability, GreenVoice highlights models that are both effective and environmentally responsible.

indicates greater difficulty for the ADD model in distinguishing real and cloned samples, signifying higher realism of the cloned voices. This results in the following set of EER values:

$$\mathrm{EER}_{\mathrm{ADD}} = \{\mathrm{EER}^i_{\mathrm{ADD}}\}^T_{i=1}.$$

From hereon, we refer to $\mathrm{EER}^i_{\mathrm{ADD}}$ as the *Realism Score* of generation model $i$.

**Similarity Test:** The similarity test measures how closely the cloned voices resemble the bonafide voices of the target speakers. This assessment uses speaker verification (SV) models, which compare two audio recordings to determine whether they belong to the same speaker. In this setup, positive trial pairs consist of two bonafide recordings of the same speaker, while negative trial pairs include one bonafide and one cloned recording. For each target model $t_i$, a test set with $v$ trial pairs (balanced between positive and negative) is evaluated using an SV model, yielding $\mathrm{EER}^i_{\mathrm{SV}}$. Higher EER values indicate greater similarity between cloned and bonafide voices. The resulting set of EER values is:

$$\mathrm{EER}_{\mathrm{SV}} = \{\mathrm{EER}^i_{\mathrm{SV}}\}^T_{i=1}$$

From hereon, we refer to $\mathrm{EER}^i_{\mathrm{SV}}$ as the *Similarity Score* of generation model $i$.

**Environmental Impact Assessment:** To complement performance-based tests, GreenVoice measures the environmental cost of synthetic speech generation. For each target model $t_i$, the carbon emissions generated from producing $c$ synthetic samples are recorded, resulting in:

$$\mathrm{E}_{\mathrm{CO2}} = \{\mathrm{E}^i_{\mathrm{CO2}}\}^T_{i=1}$$

**Max Normalization:** Since $\widehat{\mathrm{EER}}^i_{\mathrm{ADD}}$, $\mathrm{EER}^i_{\mathrm{SV}}$, and $\mathrm{E}^i_{\mathrm{CO2}}$ represent distinct metrics, max normalization is applied to make them comparable. Each score is divided by the maximum value across all target models:

$$\widehat{\mathrm{EER}}^i_{\mathrm{ADD}} = \frac{\mathrm{EER}^i_{\mathrm{ADD}}}{\max\limits_{j=1,\ldots,T} \mathrm{EER}^j_{\mathrm{ADD}}} \tag{7.1}$$

$$\widehat{\mathrm{EER}}^i_{\mathrm{SV}} = \frac{\mathrm{EER}^i_{\mathrm{SV}}}{\max\limits_{j=1,\ldots,T} \mathrm{EER}^j_{\mathrm{SV}}} \tag{7.2}$$

$$\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}} = \frac{\mathrm{E}^{i}_{\mathrm{CO2}}}{\max\limits_{j=1,\ldots,T} \mathrm{E}^{j}_{\mathrm{CO2}}} \tag{7.3}$$

These max-normalized scores, now in the range [0,1], are directly comparable and suitable for further evaluation. Similarly, for each target model, we get a max-normalized score, as represented below:

$\widehat{\mathrm{EER}}_{\mathrm{ADD}} = \{\widehat{\mathrm{EER}}^{i}_{\mathrm{ADD}}\}_{i=1}^{T}$

$\widehat{\mathrm{EER}}_{\mathrm{SV}} = \{\widehat{\mathrm{EER}}^{i}_{\mathrm{SV}}\}_{i=1}^{T}$

$\widehat{\mathrm{E}}_{\mathrm{CO2}} = \{\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}}\}_{i=1}^{T}$

**Cloning Quality Assessment:** The CQA score of a target model $t_i$ is the weighted combination of max-normalized realism and similarity scores:

$$\mathrm{CQA}_i = \frac{w_{\mathrm{ADD}} \times \widehat{\mathrm{EER}}^{i}_{\mathrm{ADD}} + w_{\mathrm{SV}} \times \widehat{\mathrm{EER}}^{i}_{\mathrm{SV}}}{w_{\mathrm{ADD}} + w_{\mathrm{SV}}} \tag{7.4}$$

where $w_{\mathrm{ADD}}$ and $w_{\mathrm{SV}}$ are the weights for realism and similarity tests, respectively.

**GreenVoice Score:** The $CQA_i$ captures model performance, while $\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}}$ quantifies environmental cost. These are combined to compute the GreenVoice Score:

$$G\text{-}Score_i = \frac{\mathrm{CQA}_i}{\left(\widehat{\mathrm{E}}^{i}_{\mathrm{CO2}}\right)^{\frac{1}{n}}} \tag{7.5}$$

Here, the scaling factor $n$ controls the influence of carbon emissions; lower $n$ increases the gap between high-cost and low-cost models.

**GreenVoice Evaluation:** Target models with high Realism Scores can be leveraged for ADD dataset creation, whereas those with high Similarity Scores can be explored for anti-spoofing dataset creation. The CQA allows selection of target models that attain the required balance in Realism and Similarity dimensions, based on the task-specific needs. Each target model's $G\text{-}Score_i$ reflects its ability to generate high-quality voice clones with minimal environmental impact. Models with higher $G\text{-}Score$ are considered superior, promoting sustainable and high-performing voice cloning models for large-scale ADD and anti-spoofing dataset creation.

## 7.4 Experimental Setup

We evaluate eight publicly available state-of-the-art (SOTA) voice cloning models, comprising five TTS models and three VC models. The TTS models are MetaVoice-1B, Coqui XTTS-v2, SV2TTS, OpenVoice-V2, and YourTTS [103, 68, 21, 132, 20], while the VC models include FreeVC, DiffVC, and SeedVC [90, 127, 93].

### 7.4.1 Realism and Similarity Tests

We conducted realism tests using two publicly available state-of-the-art (SOTA) audio deepfake detection (ADD) models: Aasist and RawNet-2 [72, 161][1]. To ensure robust and fair realism assessment, we averaged the Equal Error Rate (EER) scores obtained from both ADD models, denoted as $\mathrm{EER}_{\mathrm{ADD}}$.

For similarity tests, we used two publicly available SOTA speaker verification (SV) models: ECAPA-TDNN[2] and ResNet-TDNN[3] [135, 39, 175]. Similarly, we averaged the EER scores from both SV models to obtain a fair measure of similarity performance ($\mathrm{EER}_{\mathrm{SV}}$).

### 7.4.2 Dataset and Accent Diversity

The target TTS and VC models were evaluated on four publicly available datasets: LibriTTS (test-clean subset), CEABI (Open Source Multi-Speaker Corpora of English Accents in the British Isles), NISP, and AESRC 2020 [199, 77, 37, 148]. The LibriTTS test-clean subset primarily contains British and American English recordings, which align with the accents the target models were trained on. Thus, we consider it an *in-domain* test set.

From CEABI, we extracted recordings for four English accents—Southern, Northern, Scottish, and Welsh—which differ from standard British or American English but remain within the anglophone spectrum. These are categorized as *moderately out-of-*

[1] Aasist:`https://github.com/clovaai/aasist`
RawNet-2:`https://github.com/asvspoof-challenge/2021/tree/main/DF/Baseline-RawNet2`
[2] `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`
[3] `https://huggingface.co/speechbrain/spkrec-resnet-voxceleb`

*domain (Mod. OOD)*. NISP and AESRC provide Indian-English and Chinese-English recordings, respectively, which exhibit distinct prosodic patterns compared to British or American English. Therefore, they are labeled as *strongly out-of-domain (Str. OOD)*. Using these datasets, we evaluate seven English accents in total: *British/American, Southern, Northern, Scottish, Welsh, Indian, and Chinese*, to assess the inclusivity of the target models through GreenVoice.

### 7.4.3 Test Set Creation

Bonafide recordings from the above datasets were used to generate cloned audios with the target TTS and VC models. For each accent, the same set of speakers was used across all model test sets to ensure fair benchmarking. While creating VC-based cloned recordings, the gender of source and target speakers was kept consistent.

**Similarity Test Sets:** For every target model and English accent, we generated similarity test sets containing 20,000 trial pairs, with an equal split of positive and negative pairs.

**Realism Test Sets:** Similarly, realism test sets were generated for each target model and accent, comprising 4,000 audio clips: 2,000 bonafide and 2,000 cloned. Due to fewer male recordings in LibriTTS, its ADD test set contains 3,800 clips.

## 7.5 Experiments and Results

To demonstrate the application of GreenVoice, we evaluate eight state-of-the-art (TTS/VC) models across seven English accents, encompassing both in-domain and out-of-domain (OOD) variations. The evaluation covers four major dimensions: realism (authenticity) testing, similarity (mimicry) assessment, cost analysis, and the overall GreenVoice benchmarking.

### 7.5.1 Realism Test

A higher Realism Score for a given generation model indicates that audio deepfake detection (ADD) systems find it difficult to differentiate between bonafide speech and

the synthetic outputs generated by that model. Consequently, such a model can be utilized to create challenging, high-quality datasets for training ADD systems. Therefore, this section compares the target (TTS/VC) models based on their Realism Scores ($\mathrm{EER}_{\mathrm{ADD}}$)[4].

**Setup:** For this experiment, we created separate audio deepfake detection (ADD) test sets for each of the seven target English accents described in Section 7.4.2 and prepared as outlined in Section 7.4.3. Additionally, to account for gender, we constructed separate ADD test sets for female and male voices from LibriTTS, denoted as LibriTTS (F) and LibriTTS (M).

**Observations:** Table 7.1 presents the averaged $\mathrm{EER}_{\mathrm{ADD}}$ scores obtained using the Aasist and RawNet-2 models. We find that in-domain TTS test sets yield significantly higher EER values compared to out-of-domain (OOD) TTS test sets. A higher EER indicates that the ADD models struggled to differentiate between bonafide (real) and synthetic (cloned) audios. This implies that the target voice cloning models produce more realistic synthetic voices for in-domain accents than for OOD accents, revealing a pronounced domain-specific (accent) bias in TTS models.

In contrast, VC models exhibit high EER values across both in-domain and OOD test sets, suggesting greater robustness to accent variations. Among the VC models, FreeVC achieves the highest EER scores for most accents, indicating superior resilience to accent variability, likely due to its bottleneck architecture that disentangles content and speaker information for improved content transfer. Additionally, EER values for LibriTTS (M) test sets are slightly higher than LibriTTS (F), suggesting that VC models generate marginally more authentic male voice clones compared to female ones.

**Summary of Realism Test Findings:**

1. VC models produce more realistic synthetic voices than TTS models across diverse English accents.
2. FreeVC demonstrates the highest robustness to accent variations.
3. TTS models show a strong domain-specific (accent) bias.
4. VC models exhibit a mild gender bias, generating slightly more authentic male voice clones than female ones.

[4]Details of the realism test are provided in Section 7.3

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | **48.67** | 30.38 | 31.27 | **48.95** | **32.55** | 34.62 | 47.52 | 42.50 |
| UK/US | LibriTTS-M | In-Domain | 41.42 | **35.71** | **31.89** | 45.86 | 28.68 | **43.68** | **50.76** | **43.34** |
| UK/US | LibriTTS | In Domain | 45.92 | 33.21 | 32.36 | 47.91 | 32.15 | 39.38 | **49.70** | 44.16 |
| Northern | CEABI | Mod. OOD | 5.95 | 1.58 | 1.85 | 2.97 | 1.20 | **62.67** | 45.75 | 47.50 |
| Southern | CEABI | Mod. OOD | 5.93 | 1.48 | 2.12 | 4.17 | 1.22 | **57.40** | 43.03 | 49.01 |
| Scottish | CEABI | Mod. OOD | 6.03 | 1.53 | 2.00 | 3.53 | 1.32 | **62.15** | 45.36 | 47.695 |
| Welsh | CEABI | Mod. OOD | 18.23 | 6.72 | 7.88 | 12.23 | 7.05 | **75.15** | 58.43 | 52.10 |
| Indian | NISP | Str. OOD | 45.95 | 9.78 | 9.50 | 8.40 | 2.97 | 44.88 | 45.47 | **51.08** |
| Chinese | AESRC | Str. OOD | 8.48 | 3.35 | 6.00 | 10.00 | 3.97 | **51.15** | 49.63 | 46.00 |

Table 7.1: The table presents the Realism Scores ($\mathrm{EER}_{\mathrm{ADD}}\%$) of target (TTS/VC) models computed using Aasist and RawNet-2 ADD models, without domain adaptation. Key observations: (1) VC models produce slightly more realistic male voice clones (LibriTTS-M) compared to female voice clones (LibriTTS-F). (2) TTS models generate more realistic voice clones for in-domain accents than for moderately (Mod. OOD) and strongly (Str. OOD) out-of-domain accents. (3) VC models exhibit greater robustness to accent variations than TTS models. (4) FreeVC consistently produces the most realistic voice clones among the evaluated models for most OOD accents.

### 7.5.2 Similarity Test

A higher Similarity Score for a given generation model indicates that speaker verification (SV) systems have difficulty distinguishing between synthesized voice clones and the genuine speech of the target speakers. As a result, such models can be utilized to generate challenging, high-quality anti-spoofing datasets for improving robustness against impersonation attacks. Therefore, this section compares the target (TTS/VC) models based on their Similarity Scores ($\mathrm{EER}_{\mathrm{SV}}$)[5].

**Setup:** For this experiment, we generated separate similarity test sets for each of the seven target English accents described in Section 7.4.2, following the procedure in Section 7.4.3. Additionally, we prepared distinct test sets for female and male voices from LibriTTS, referred to as LibriTTS (F) and LibriTTS (M).

**Observations:** Table 7.2 presents the averaged $\mathrm{EER}_{\mathrm{SV}}$ scores on these test sets. In-domain test sets show substantially higher EER values compared to out-of-domain (OOD) sets. A higher EER indicates that the SV models struggled to differentiate between positive and negative trial pairs. When an SV model finds it difficult to distinguish these pairs, it implies that the cloned voices closely resemble the bonafide target voices. Therefore, higher EERs for in-domain test sets suggest that target models produce more accurate voice clones for in-domain accents than for OOD accents, reflecting a strong accent-specific bias. Among the evaluated models, SeedVC produces the most similar voice clones to bonafide targets, likely due to its architecture reducing

[5]Details about the similarity test are in Section 7.3.

timbre leakage and enhancing alignment. Moreover, most models exhibit higher EERs for LibriTTS-F than LibriTTS-M, indicating that female voice clones better replicate bonafide voices than male clones.

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In Domain | **11.30** | **30.58** | **34.08** | **17.29** | **24.87** | **12.15** | **9.01** | 36.38 |
| UK/US | LibriTTS-M | In Domain | 9.57 | 29.77 | 29.27 | 16.42 | 24.04 | 11.83 | 7.40 | **37.17** |
| UK/US | LibriTTS | In Domain | 10.59 | 30.68 | 32.49 | 17.33 | 25.45 | 12.30 | 8.28 | **37.53** |
| Northern | CEABI | Mod. OOD | 1.99 | 5.42 | 8.47 | 1.24 | 3.06 | 3.37 | 1.71 | **25.33** |
| Southern | CEABI | Mod. OOD | 2.17 | 5.90 | 8.95 | 1.32 | 2.68 | 3.10 | 1.40 | **26.61** |
| Scottish | CEABI | Mod. OOD | 1.17 | 4.84 | 11.94 | 2.08 | 5.52 | 3.25 | 1.61 | **26.92** |
| Welsh | CEABI | Mod. OOD | 2.11 | 6.73 | 10.66 | 1.66 | 2.79 | 3.12 | 1.27 | **24.82** |
| Indian | NISP | Str. OOD | 2.11 | 4.31 | 4.02 | 0.61 | 0.80 | 1.95 | 0.60 | **26.33** |
| Chinese | AESRC | Str. OOD | 3.485 | 8.00 | 16.50 | 3.11 | 5.10 | 6.92 | 3.00 | **29.61** |

Table 7.2: Similarity Scores ($\mathrm{EER}_{\mathrm{SV}}\%$) of target (TTS/VC) models computed using ECAPA-TDNN and ResNet-TDNN SV models, without domain adaptation. Key observations: (1) For most models, female voice clones (LibriTTS-F) more closely resemble the bonafide target voices than male voice clones (LibriTTS-M). (2) SeedVC generates voice clones with the highest similarity to bonafide target voices among all models. (3) In-domain voice clones exhibit greater similarity to target voices compared to moderately (Mod. OOD) or strongly (Str. OOD) out-of-domain (OOD) voice clones.

**Summary of Similarity Test Findings:**

1. Target models exhibit pronounced accent-specific bias.
2. SeedVC demonstrates the greatest robustness to accent variations.
3. Female voice clones more closely mimic bonafide voices than male clones.

### 7.5.3 Cost and Sustainability Analysis

Most existing voice cloning benchmarks focus exclusively on model performance [29]. While performance evaluation is important, we highlight the significance of cost-aware benchmarking. Assessing models based on cost is essential for encouraging environmentally sustainable and energy-efficient voice cloning. Accordingly, we evaluate target (TTS/VC) models using cost-related metrics, summarized in Table 7.3.

**Setup:** In this experiment, each target model was used to generate 1,000 voice clones on an NVIDIA A6000 GPU, while recording the cloning time. Carbon emissions and electricity consumption were also tracked during inference[6].

**Observations:** As shown in Table 7.3, MetaVoice-1B has the largest number of parameters, while SV2TTS has the fewest. In terms of speed, MetaVoice-1B is the slowest, whereas YourTTS is the fastest among the target models. Additionally, YourTTS

[6]We used the CodeCarbon library [36]

produces the lowest carbon emissions and electricity usage during voice cloning, likely due to its efficient architecture. Consequently, YourTTS emerges as the most environmentally friendly target model.

| Measurement | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|
| #Parameters (M) | 84.6 | 1265 | 466.8 | **36.7** | 86.8 | 356.2 | 141.6 | 310.2 |
| Cloning Time (min) | 25.42 | 348.45 | 65.58 | 218.97 | **1.89** | 3.76 | 37.34 | 127.91 |
| Carbon Emissions (kgCO2eq) | 0.066 | 1.49 | 0.166 | 0.493 | **0.004** | 0.01 | 0.154 | 0.302 |
| Energy Consumed (kWh) | 0.093 | 2.094 | 0.233 | 0.691 | **0.006** | 0.014 | 0.217 | 0.424 |

Table 7.3: Table shows cost-based benchmarks for the target models.

### 7.5.4 GreenVoice Benchmarking

Next, we evaluate the target (TTS/VC) models using our proposed GreenVoice score ($G\text{-}Score$), encouraging both efficiency and sustainability in voice cloning models for large-scale ADD and anti-spoofing dataset creation.

**Setup:** The scaling factor is set to $n = 4$. We consider two scenarios: (A) prioritizing the realism of voice clones, and (B) prioritizing high similarity between cloned and bonafide voices. For Scenario (A), the weights in $CQA$ are $w_{\text{ADD}} = 4$ and $w_{\text{SV}} = 1$, while for Scenario (B), $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 4$.

**Observations:** Table 7.4 and Table 7.5 show the $G\text{-}Score$ results for Scenarios (A) and (B). For most target models, in-domain accents achieve higher $G\text{-}Score$ values than out-of-domain (OOD) accents, indicating more efficient cloning of in-domain accents. YourTTS attains the highest $G\text{-}Score$ for in-domain accents under both scenarios. In Scenario A, FreeVC scores highest for OOD accents, consistent with its superior performance in the realism test (Section 7.5.1) and its sustainability (Section 7.5.3). In Scenario B, SeedVC achieves the highest $G\text{-}Score$ for OOD accents, aligning with its strong performance in the similarity test (Section 7.5.2). No notable gender bias is observed in $G\text{-}Score$ values between LibriTTS-F and LibriTTS-M test sets.

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | **1.87** | 0.66 | **1.21** | **1.28** | **2.94** | 2.21 | 1.46 | **1.33** |
| UK/US | LibriTTS-M | In-Domain | 1.53 | **0.72** | 1.14 | 1.16 | 2.55 | **2.63** | **1.48** | 1.32 |
| UK/US | LibriTTS | In-Domain | 1.73 | 0.70 | 1.20 | 1.23 | **2.87** | 2.44 | 1.49 | 1.36 |
| Northern | CEABI | Mod. OOD | 0.2 | 0.06 | 0.16 | 0.11 | 0.17 | **2.89** | 1.05 | 1.2 |
| Southern | CEABI | Mod. OOD | 0.22 | 0.06 | 0.17 | 0.14 | 0.16 | **2.88** | 1.08 | 1.32 |
| Scottish | CEABI | Mod. OOD | 0.19 | 0.06 | 0.2 | 0.11 | 0.25 | **2.88** | 1.05 | 1.21 |
| Welsh | CEABI | Mod. OOD | 0.46 | 0.13 | 0.29 | 0.24 | 0.43 | **2.88** | 1.11 | 1.12 |
| Indian | NISP | Str. OOD | 1.6 | 0.19 | 0.31 | 0.22 | 0.23 | **2.51** | 1.26 | 1.49 |
| Chinese | AESRC | Str. OOD | 0.34 | 0.11 | 0.36 | 0.28 | 0.42 | **2.96** | 1.40 | 1.37 |

Table 7.4: $G\text{-}Score$ for Scenario A with $w_{\text{ADD}} = 4$ and $w_{\text{SV}} = 1$, emphasizing realism of voice clones. YourTTS achieves the highest scores for in-domain accents, while FreeVC leads for out-of-domain accents.

**Key Insights:**

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | **0.98** | **0.80** | **1.52** | **0.77** | **2.99** | 1.43 | **0.69** | 1.45 |
| UK/US | LibriTTS-M | In-Domain | 0.80 | 0.78 | 1.31 | 0.70 | 2.77 | **1.49** | 0.63 | 1.45 |
| UK/US | LibriTTS | In-Domain | 0.89 | 0.79 | 1.42 | 0.74 | **2.95** | 1.47 | 0.66 | 1.46 |
| Northern | CEABI | Mod. OOD | 0.18 | 0.18 | 0.47 | 0.06 | 0.44 | 1.07 | 0.35 | **1.42** |
| Southern | CEABI | Mod. OOD | 0.19 | 0.18 | 0.48 | 0.07 | 0.37 | 1.02 | 0.34 | **1.45** |
| Scottish | CEABI | Mod. OOD | 0.12 | 0.15 | 0.63 | 0.1 | 0.74 | 1.04 | 0.34 | **1.42** |
| Welsh | CEABI | Mod. OOD | 0.25 | 0.23 | 0.63 | 0.11 | 0.48 | 1.05 | 0.35 | **1.40** |
| Indian | NISP | Str. OOD | 0.53 | 0.17 | 0.28 | 0.07 | 0.16 | 0.82 | 0.35 | **1.49** |
| Chinese | AESRC | Str. OOD | 0.28 | 0.23 | 0.81 | 0.16 | 0.67 | 1.35 | 0.49 | **1.46** |

Table 7.5: $G\text{-}Score$ for Scenario B with $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 4$, emphasizing similarity test performance. YourTTS achieves the highest scores for in-domain accents, while SeedVC leads for out-of-domain accents.

1. *FreeVC* achieves the highest score in realism tests across most accents.
2. *SeedVC* attains the highest score in similarity tests across accents.
3. *YourTTS* produces the lowest carbon emissions among the target (TTS/VC) models.
4. When prioritizing realism of synthetic voices (Scenario A), *FreeVC* is a sustainable option.
5. When prioritizing similarity between bonafide and cloned voices (Scenario B), *SeedVC* is the sustainable choice.

## 7.6 GreenVoice Score: Scaling Factor Impact

Tables 7.6, 7.7, and 7.8 present $G\text{-}Score$ values for scaling factors $n = 2, 4,$ and $8$, computed with $w_{\text{ADD}} = w_{\text{SV}} = 1$. In each table, the highest and lowest $G\text{-}Score$ values in every row are highlighted. We observe that smaller $n$ values lead to larger variations in $G\text{-}Score$. For example, at $n = 2$, the difference between the highest and lowest average $G\text{-}Score$ is 6.46 (6.80–0.34), whereas at $n = 4$ it reduces to 1.61 (1.95–0.34), and at $n = 8$ it further decreases to 0.78 (1.12–0.34). This indicates that increasing $n$ diminishes the influence of carbon emissions on the $G\text{-}Score$, while a lower $n$ amplifies it. Additionally, higher $n$ generally lowers $G\text{-}Score$ values overall, except for MetaVoice-1B, whose $G\text{-}Score$ remains unchanged due to its extremely high recorded carbon emissions.

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | 3.1 | **0.73** | 2.36 | 1.28 | **13.01** | 6.35 | 1.9 | 2.07 |
| UK/US | LibriTTS-M | In-Domain | 2.55 | **0.75** | 2.12 | 1.17 | **11.69** | 7.19 | 1.86 | 2.06 |
| UK/US | LibriTTS | In-Domain | 2.87 | **0.74** | 2.27 | 1.24 | **12.79** | 6.84 | 1.9 | 2.1 |
| Northern | CEABI | Mod. OOD | 0.41 | 0.12 | 0.55 | 0.08 | 1.35 | **6.92** | 1.24 | 1.95 |
| Southern | CEABI | Mod. OOD | 0.44 | 0.12 | 0.56 | 0.11 | 1.18 | **6.81** | 1.25 | 2.06 |
| Scottish | CEABI | Mod. OOD | 0.33 | 0.1 | 0.71 | 0.12 | 2.18 | **6.84** | 1.23 | 1.96 |
| Welsh | CEABI | Mod. OOD | 0.78 | 0.18 | 0.8 | 0.2 | 2.00 | **6.87** | 1.29 | 1.88 |
| Indian | NISP | Str. OOD | 2.33 | 0.18 | 0.51 | 0.16 | 0.85 | **5.81** | 1.42 | 2.22 |
| Chinese | AESRC | Str. OOD | 0.67 | 0.17 | 1.01 | 0.26 | 2.41 | **7.53** | 1.67 | 2.11 |
| **Average** | | | 1.5 | 0.34 | 1.21 | 0.51 | 5.27 | **6.80** | 1.53 | 2.05 |

Table 7.6: $G\text{-}Score$ when the scaling factor $n$ is set to 2. Here, $w_{\text{ADD}}$ = 1 and $w_{\text{SV}}$ = 1

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | 1.42 | **0.73** | 1.36 | 0.97 | **2.96** | 1.82 | 1.07 | 1.39 |
| UK/US | LibriTTS-M | In-Domain | 1.17 | **0.75** | 1.23 | 0.89 | **2.66** | 2.06 | 1.06 | 1.38 |
| UK/US | LibriTTS | In-Domain | 1.31 | **0.74** | 1.31 | 0.94 | **2.91** | 1.96 | 1.08 | 1.41 |
| Northern | CEABI | Mod. OOD | 0.19 | 0.12 | 0.31 | **0.06** | 0.31 | **1.98** | 0.70 | 1.31 |
| Southern | CEABI | Mod. OOD | 0.20 | 0.12 | 0.32 | **0.08** | 0.27 | **1.95** | 0.71 | 1.38 |
| Scottish | CEABI | Mod. OOD | 0.15 | 0.10 | 0.41 | **0.09** | 0.50 | **1.96** | 0.70 | 1.32 |
| Welsh | CEABI | Mod. OOD | 0.36 | 0.18 | 0.46 | **0.15** | 0.45 | **1.97** | 0.73 | 1.26 |
| Indian | NISP | Str. OOD | 1.07 | 0.18 | 0.29 | **0.12** | 0.19 | **1.66** | 0.81 | 1.49 |
| Chinese | AESRC | Str. OOD | 0.31 | **0.17** | 0.58 | 0.20 | 0.55 | **2.16** | 0.94 | 1.42 |
| **Average** | | | 0.69 | **0.34** | 0.70 | 0.39 | 1.20 | **1.95** | 0.87 | 1.37 |

Table 7.7: $G\text{-}Score$ when the scaling factor $n$ is set to 4. Here, $w_{\text{ADD}}$ = 1 and $w_{\text{SV}}$ = 1.

| Accent | Dataset | Category | OpenVoice-V2 | MetaVoice-1B | XTTS-v2 | SV2TTS | YourTTS | FreeVC | DiffVC | SeedVC |
|---|---|---|---|---|---|---|---|---|---|---|
| UK/US | LibriTTS-F | In-Domain | 0.96 | **0.73** | 1.04 | 0.85 | **1.41** | 0.97 | 0.81 | 1.14 |
| UK/US | LibriTTS-M | In-Domain | 0.79 | **0.75** | 0.93 | 0.77 | **1.27** | 1.1 | 0.8 | 1.13 |
| UK/US | LibriTTS | In-Domain | 0.89 | **0.74** | 1 | 0.82 | **1.39** | 1.05 | 0.81 | 1.15 |
| Northern | CEABI | Mod. OOD | 0.13 | 0.12 | 0.24 | **0.06** | 0.15 | 1.06 | 0.53 | **1.07** |
| Southern | CEABI | Mod. OOD | 0.14 | 0.12 | 0.25 | **0.07** | 0.13 | 1.04 | 0.53 | **1.13** |
| Scottish | CEABI | Mod. OOD | 0.1 | 0.1 | 0.31 | **0.08** | 0.24 | 1.05 | **0.52** | 1.08 |
| Welsh | CEABI | Mod. OOD | 0.24 | 0.18 | 0.35 | **0.13** | 0.22 | **1.05** | 0.55 | 1.03 |
| Indian | NISP | Str. OOD | 0.72 | 0.18 | 0.22 | **0.11** | 0.09 | 0.89 | 0.61 | **1.22** |
| Chinese | AESRC | Str. OOD | **0.21** | 0.17 | 0.44 | 0.17 | 0.26 | 1.15 | 0.71 | **1.16** |
| **Average** | | | 0.46 | **0.34** | 0.53 | 0.34 | 0.57 | 1.04 | 0.65 | **1.12** |

Table 7.8: $G\text{-}Score$ when the scaling factor $n$ is set to 8. Here, $w_{\text{ADD}}$ = 1 and $w_{\text{SV}}$ = 1.

## 7.7 Social Impact

GreenVoice benchmarks existing publicly available models to highlight those that are both effective and sustainable for social good as follows:

1. **Assisting Responsible AI Researchers:** Through GreenVoice, responsible AI researchers can easily identify high-performing and low-cost publicly available voice cloning models that may be misused for audio deepfake related fraud. Consequently, they may generate synthetic data using these cloning models and train audio deepfake detection and anti-spoofing systems to enhance robustness of these security systems against fake audio generated by these cloning models, thus combating deepfake-related fraud.

2. **Social Good Applications:** High-performing and sustainable models identified through GreenVoice have socially beneficial applications, such as assistive technologies for individuals with speech disorders, education, and entertainment.

3. **Environmental Impact:** About 90% of ACL papers, 80% of NeurIPS papers and 75% of CVPR papers target accuracy, instead of efficiency for evaluation [158]. Similarly, training a BERT model on a single GPU causes carbon emissions that are equivalent to a trans-American flight [158]. Yet, environmental assessments are uncommon in speech processing. The widespread use of voice cloning models with millions or billions of parameters without such studies is concerning from ethical and sustainability standpoints. GreenVoice explicitly promotes inference-efficient models with lower carbon emissions.

In this way, GreenVoice promotes responsible usage by highlighting sustainable voice cloning models for socially beneficial applications, while also enabling responsible AI researchers to advance defenses against their malicious use.

## 7.8 Conclusions and Future Work

This chapter presented GreenVoice, an automated, environmentally aware benchmarking framework for text-to-speech (TTS) and voice conversion (VC) models. GreenVoice provides a comprehensive performance-cost evaluation through the G-Score, which identifies efficient models as those with higher scores. The G-Score combines Cloning Quality Assessment (CQA) and Environmental Impact Assessment (EIA). CQA evaluates the quality of generated voice clones using realism and similarity tests, while EIA measures the carbon emissions associated with voice cloning. The realism test assesses the authenticity of synthetic voices, whereas the similarity test measures how closely the clones mimic bonafide target voices.

Using GreenVoice, eight cloning models were benchmarked across seven English accents. The realism tests highlighted a pronounced accent-specific bias in TTS models and a strong robustness of VC models to out-of-domain (OOD) accents. Similarity tests also revealed a significant accent-specific bias among target models. The EIA evaluation showed that YourTTS is the most environmentally sustainable model. Overall, GreenVoice indicates that YourTTS performs best for in-domain cloning, while FreeVC and SeedVC are preferable for OOD scenarios. Specifically, FreeVC excels when synthetic voice realism is prioritized, whereas SeedVC is optimal when achieving high similarity to bonafide voices is the goal.

This study highlights several promising directions for future research. First, GreenVoice facilitates the cost-efficient identification of sustainable generation models capable of producing challenging synthetic speech in target languages and accents. The data generated using these models can therefore be used to construct high-quality multilingual datasets for audio deepfake detection and anti-spoofing research. Moreover, GreenVoice provides a framework to analyze the behavior of generation models across diverse linguistic settings. Such analyses can drive more comprehensive investigations into model architectures and training methodologies. The resulting insights can contribute to the development of more robust countermeasures for detecting synthetic audio generated by these models. Furthermore, large-scale synthetic speech generation incurs a high inference cost.

## 7.9 Limitations

While GreenVoice represents an initial step toward automated, environmentally aware benchmarking, several aspects could be improved:

1. **Weight assignments:** The parameters $n$, $w_{\text{ADD}}$, and $w_{\text{SV}}$ should be selected based on the specific benchmarking goals and user priorities. In Section 7.5.4, we set $n = 4$, but also explored $n = 2$ and $n = 8$ (Section 7.6). Similarly, we used $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 4$ for Scenario (A), and $w_{\text{ADD}} = 4$ and $w_{\text{SV}} = 1$ for Scenario (B), with additional experiments using $w_{\text{ADD}} = 1$ and $w_{\text{SV}} = 1$ reported in the Section 7.6 .

2. **Dependence on external tools:** GreenVoice relies on audio deepfake detection (ADD) and speaker verification (SV) models for assessing cloning quality. Consequently, any limitations of these models can affect GreenVoice results. Similarly, carbon emission estimates depend on GPU runtime, hardware, and region-specific carbon intensity, which may vary. To mitigate variability, we averaged results from two state-of-the-art ADD and SV models and ran all inferences on the same server, GPU, and day. During carbon tracking, the server was dedicated exclusively to one model at a time.

3. **Scope:** Currently, GreenVoice supports only TTS and VC models. However, the framework can be extended to other model types, such as automatic speech recognition (ASR). In this work, we benchmarked five TTS and three VC models across seven English accents, but future work could include additional models, accents, and languages. Our carbon emission analysis focuses on inference-time emissions, excluding training, long-term deployment, or hardware lifecycle contributions. While including these factors would provide a fuller environmental assessment, limiting the scope to inference allows for reproducible, efficient, and standardized comparisons. Inference workloads are easier to replicate and less resource-intensive than training, enabling fair benchmarking under controlled conditions.

Overall, we hope that GreenVoice encourages researchers to adopt Green AI practices, fostering environmentally responsible development in speech technologies.

## 7.10 Ethical Considerations

Synthetic speech has a wide range of applications, including entertainment, personalized digital assistants, and audiobook narration. At the same time, it poses potential risks, such as the creation of audio deepfakes for spreading misinformation. To address these concerns, we developed GreenVoice to encourage the responsible use of voice cloning models. Although there is a possibility that malicious users could exploit

GreenVoice to identify highly efficient voice cloning models, GreenVoice is designed to primarily serve Responsible AI researchers. Responsible AI researchers can leverage GreenVoice to highlight efficient models and subsequently enhance audio deepfake detection and anti-spoofing technologies, mitigating potential misuse. In addition, GreenVoice encourages the adoption of Green AI principles by reporting the environmental impact of large voice cloning models, addressing a notable gap in current research. Overall, we strongly advocate for the use of GreenVoice to support socially beneficial and ethical applications.

# CHAPTER 8

# Conclusions and Future Research Directions

This thesis investigates synthetic speech in multilingual, low-resource settings, with a focus on security, fairness, and sustainability. It integrates dataset creation, machine and human evaluation, cost-efficient linguistic bias mitigation in speaker verification, and an automated, environment-aware framework for evaluating synthetic speech generation models. Collectively, these contributions advance responsible, inclusive, and sustainable speech processing in diverse linguistic contexts.

## 8.1 Key Contributions and Findings

### 8.1.1 IndicSynth: Multilingual Synthetic Speech for Audio Deepfake Detection and Anti-Spoofing

Chapter 2 highlighted the risks posed by synthetic speech, including its misuse for misinformation and impersonation. Consequently, there is an urgent need for robust audio deepfake detection (ADD) and anti-spoofing systems. Existing ADD and anti-spoofing systems often exhibit language dependency. However, the lack of publicly available multilingual synthetic speech datasets hinders multilingual ADD and anti-spoofing research. The problem intensifies in linguistically diverse countries, such as India. To address this gap, IndicSynth was introduced—a large-scale synthetic speech dataset comprising approximately 4,000 hours of audio from 989 target speakers (456 female, 533 male) across 12 low-resource Indian languages. IndicSynth is organized into mimicry and diversity subsets to balance the need for realistic voice imitation and broad synthetic diversity. Evaluation using state-of-the-art ADD and speaker verification (SV) models demonstrated that existing systems struggle with linguistic biases. Additionally, SV models are vulnerable to impersonation attacks, emphasizing the critical utility of IndicSynth. However, human perception of IndicSynth audios is essential for a thorough evaluation of the dataset before using it for multilingual ADD and anti-spoofing research.

### 8.1.2 Understanding Human Perception of Synthetic Speech

Chapter 3 complemented the machine-based evaluations by investigating human perception of IndicSynth audios. A user study with 93 participants was conducted to assess the naturalness of the audios and to evaluate how closely the mimicry subset resembles the corresponding bonafide target voices, thereby examining the quality of both the mimicry and diversity subsets. The study consisted of three tasks: (1) classifying test audios as real or synthetic, (2) identifying a bonafide audio from a set of real and synthetic recordings, and (3) rating the similarity of synthetic audios in the mimicry subset to their corresponding bonafide target voices. The results revealed that participants often struggled to reliably distinguish synthetic speech from real recordings, and the perceived naturalness of synthetic audios was comparable to that of real speech. These findings highlight the potential risks associated with high-quality synthetic speech while emphasizing the importance of datasets like IndicSynth for supporting research on audio deepfake detection and anti-spoofing.

### 8.1.3 Cross-Task Dataset Profiling

Chapter 4 introduced Task-Lens, a cross-task utility-based profiling framework designed to evaluate the applicability of speech datasets across multiple downstream tasks. Applying Task-Lens to 34 publicly available Indian speech datasets spanning 26 languages revealed critical gaps in resources for audio deepfake detection, speech emotion recognition, and TTS tasks, particularly in underrepresented languages. IndicSynth ranked among the top datasets for cross-task utility, confirming its broad relevance. Task-Lens also provides a systematic methodology to guide future dataset creation and selection, enabling more inclusive and task-agnostic research in multilingual speech processing.

### 8.1.4 Linguistic Bias Mitigation in Speaker Verification

Chapters 5 and 6 focused on mitigating linguistic biases in speaker verification (SV) systems. Insights from Chapter 2 revealed that SV models are susceptible to linguistic biases, while Chapter 4 highlighted the scarcity of data for underrepresented Indian lan-

guages. Motivated by these findings, Chapter 5 introduced FAtNet, a cost-efficient approach to reducing fully cross-lingual bias in SV. FAtNet leverages lightweight frame-level embeddings and attention mechanisms to emphasize discriminative features, improving performance on out-of-domain languages without requiring domain adaptation.

While FAtNet addressed fully cross-lingual bias, Chapter 6 extended this work to partially cross-lingual scenarios, where one recording in a trial pair may belong to the training language and the other to an unseen language. EcoSpeak, introduced in Chapter 6, incorporates contrastive linguistic (CL) attention and bias-correcting mechanisms to selectively emphasize or suppress parts of speaker embeddings based on linguistic differences. Beyond providing cost-efficient strategies for linguistic bias mitigation, Chapter 6 also promoted Green AI practices by reporting carbon emissions for both baseline models and EcoSpeak, highlighting environmentally responsible approaches in speech processing.

### 8.1.5 Responsible and Environment-Aware Voice Cloning

Chapter 7 extended the Green AI perspective from Chapter 6 by focusing on voice cloning models, which are more computationally intensive, contain millions of parameters, and consequently have higher carbon emissions. While earlier chapters highlighted the risks of synthetic speech, including misuse for deepfakes and impersonation attacks, this chapter emphasized the potential of synthetic speech for socially beneficial applications and the need for responsible usage. Chapter 7 presented GreenVoice, an automated comparative evaluation framework for synthetic speech generation models. GreenVoice overcomes the limitations of human evaluations and is used to evaluate eight voice cloning models across seven English accents. By combining cloning quality assessment with carbon emissions evaluation, GreenVoice highlights high-performing and sustainable voice cloning models for large-scale creation of multilingual audio deepfake detection and anti-spoofing datasets, thus supporting inclusive and responsible speech processing.

## 8.2 Research Narrative

First, Chapter 2 introduced IndicSynth, a large-scale synthetic speech dataset to facilitate multilingual audio deepfake detection and anti-spoofing research. The chapter demonstrated the vulnerability of existing audio deepfake detection and speaker verification models against multilingual synthetic speech attacks. Chapter 3 further presented a human evaluation of IndicSynth, demonstrating the quality of IndicSynth audios and the limited human ability to identify native-language-based audio deepfakes. However, the human evaluation was limited by the availability of fewer participants for some of the low-resource languages, highlighting a potential challenge faced by researchers in such resource-creation initiatives. Furthermore, to explore the broader utility of IndicSynth and other existing Indian speech resources for more downstream tasks, Chapter 4 introduced Task-Lens, a cross-task utility-based profiling framework. The chapter suggested a broader utility of IndicSynth for more downstream tasks based on the available metadata.

While Chapter 2 revealed the vulnerability of ADD and speaker verification models against multilingual synthetic speech attacks, these security-sensitive models are influenced by both linguistic variations in synthetic speech and the deviations in synthetic speech generation methods. Given the large number of existing languages and generation models, simultaneously handling both these biases is complex. Therefore, Chapter 5 and Chapter 6 investigate cost-effective solutions to mitigate linguistic biases in speaker verification models. In addition to computational cost, Chapter 6 discusses the environmental impact of proposed solutions. Chapter 7 extends the idea of Green speech processing in the context of large-scale dataset creation for audio deepfake detection and anti-spoofing research. The chapter introduces an automated comparative evaluation framework for synthetic speech generation models to cost-effectively highlight high-performing and sustainable generation models. Thus, by mitigating human evaluation challenges in synthetic speech dataset creation, the automated evaluation framework enhances broader participation in resource creation initiatives across diverse languages and accents. At the same time, the framework minimizes the environmental impact by promoting sustainable generation models for large-scale dataset creation.

## 8.3 Overall Inferences

Across the thesis, several recurring insights emerge:

1. **Synthetic speech poses both risks and opportunities.** High-quality synthetic voices can deceive humans and SV systems but also enable accessibility, entertainment, and personalized technologies.
2. **Linguistic and accent biases limit system fairness.** Effective mitigation requires cost-efficient, data-driven, and environmentally conscious approaches.
3. **Resource scarcity remains a major challenge.** Multilingual datasets like IndicSynth and frameworks like Task-Lens are critical for guiding research and dataset creation.
4. **Responsible AI and Green AI principles are increasingly important.** Quantifying environmental impact and ethical considerations ensures sustainable and socially beneficial speech technologies.

## 8.4 Limitations

Despite addressing key challenges in multilingual synthetic speech, several limitations remain:

1. IndicSynth currently covers 12 Indian languages; expanding to additional languages, voices, and synthesis models would further improve generalization.
2. User studies are limited in scope and participant diversity, reflecting the inherent challenges of conducting large-scale human evaluations for low-resource languages. While GreenVoice provides an automated and scalable alternative to human evaluation, further large-scale and diverse human studies are necessary for comprehensive validation.
3. Task-Lens currently assesses the potential utility of existing datasets for downstream tasks solely based on available metadata, without empirical validation through downstream model training. For example, training models such as automatic speech recognition (ASR) on datasets identified as suitable by Task-Lens could provide stronger validation of its effectiveness.
4. Bias mitigation strategies in FAtNet, EcoSpeak, and GreenVoice require further validation across a broader range of datasets, languages, and accents. Similarly, experiments across the thesis are primarily limited to languages and accents; extending the analysis to other demographic attributes, such as gender and age, would provide a more comprehensive evaluation.
5. Environmental impact analysis primarily focuses on inference, excluding the effects of model training and long-term deployment.

## 8.5 Future Directions

Building on the contributions of this thesis, several promising directions for future research are:

1. **Dataset Expansion and Multilingual Benchmarking:**
   - Extend IndicSynth to additional low-resource languages and include more voice cloning models.
   - Incorporate multilingual and code-switched scenarios to better capture real-world challenges.

2. **Human-Centered Evaluations:**
   - Conduct larger and more diverse user studies to evaluate perceptual naturalness and mimicry across languages and accents.
   - Investigate longitudinal adaptation of listeners to synthetic speech to improve human-centered detection strategies.

3. **Advanced Bias Mitigation Techniques:**
   - Extend bias analysis and mitigation strategies to other speech processing tasks, such as automatic speech recognition, emotion recognition, audio deepfake detection, and text-to-speech synthesis.
   - Work towards developing multitask, domain-invariant architectures and integrating linguistic, accent, and socio-demographic features to enable scalable, cross-task solutions for bias mitigation in multilingual and low-resource settings.

4. **Environmentally Responsible Speech Technologies:**
   - Extend GreenVoice to multilingual, multi-task evaluation, including automatic speech recognition, emotion recognition, and text-to-speech systems.
   - Track the full life-cycle environmental impact, including training and deployment, to optimize Green AI practices.

5. **Responsible Synthetic Speech Applications:**
   - Encourage socially beneficial uses of voice cloning, such as accessibility tools, language preservation, and education.
   - Investigate defenses against misuse, including audio deepfakes, impersonation, and misinformation, leveraging robust datasets and automated benchmarking frameworks.

## 8.6 Concluding Remarks

This thesis provides a coherent narrative spanning multilingual synthetic speech dataset creation, human perception analysis, cross-task profiling, linguistic bias mitigation, and

an automated environment-aware evaluation framework for synthetic speech generation models.

First, this thesis introduced IndicSynth, a 4,000-hour multilingual synthetic speech dataset spanning 12 low-resource Indian languages. The dataset addresses an important resource gap in multilingual audio deepfake detection and anti-spoofing research. Experimental evaluations on IndicSynth further demonstrated the vulnerability of existing audio deepfake detection and speaker verification models to multilingual synthetic speech attacks, underscoring the need for more robust and inclusive defenses. The dataset provides rich metadata, including gender information, target speaker identifiers, and text transcripts. Given the availability of this metadata, IndicSynth may also be explored in the future for additional downstream tasks, such as automatic speech recognition, thereby helping to mitigate task-resource gaps for underrepresented languages.

Second, the thesis examined human ability to identify native-language audio deepfakes using audio samples from IndicSynth. The empirical findings showed that participants had limited ability to reliably distinguish synthetic speech from bonafide speech, with performance approaching random guessing. These results further emphasize the risks posed by realistic synthetic speech in low-resource multilingual settings, while also providing human-centered evidence of the quality of IndicSynth.

Third, the thesis presented Task-Lens, a cross-task profiling framework designed to highlight the potential utility of existing multilingual speech datasets across diverse downstream tasks beyond their originally intended scope. The framework performs this profiling based on the availability of metadata in existing datasets that are relevant to supporting different downstream tasks. Using Task-Lens, this thesis profiled 34 Indian speech datasets, including IndicSynth, across 26 languages and eight downstream tasks. This profiling helps mitigate task-resource gaps by providing practical guidance for researchers working on underrepresented languages and low-resource speech tasks.

Fourth, the thesis builds on the observation that speaker verification models are vulnerable to multilingual audio spoofing attacks. Since multilingual audio spoofing is a complex challenge shaped by both linguistic bias and synthetic generation model-specific biases, this thesis focuses specifically on the linguistic bias component. To address this aspect, the thesis proposed FAtNet and EcoSpeak, two cost-efficient approaches for mitigating linguistic bias in speaker verification systems. EcoSpeak further

broadens this contribution by bringing sustainability into the discussion and introducing the perspective of Green speech processing in speaker verification. Overall, these contributions provide effective strategies that can support the future development of more language-invariant and robust security systems.

Fifth, the thesis concludes by proposing GreenVoice, an automated environment-aware evaluation framework designed to address common challenges in the large-scale creation of audio deepfake detection and anti-spoofing datasets. The framework introduces novel metrics for evaluating speech quality based on the realism of synthetic speech and the similarity between cloned synthetic voices and their corresponding bonafide target voices, along with a unified cloning quality assessment score. Furthermore, GreenVoice takes into account the environmental impact of large-scale synthetic speech generation and introduces the novel G-Score. G-Score highlights high-performing and sustainable generation models for large-scale audio deepfake detection and anti-spoofing dataset creation.

GreenVoice offers three key benefits. First, it enhances inclusivity in dataset creation by helping resource-constrained researchers identify high-performing synthetic speech generation models that can be used across languages and accents without fine-tuning. Second, its automated evaluation complements traditional human evaluation of synthetic speech, which is costly, difficult to scale, and less reproducible. Third, the framework explicitly accounts for the environmental cost of large-scale dataset creation by identifying sustainable generation models. Beyond dataset creation, the framework can also be used to study the behavior of synthetic speech generation models across diverse real-world scenarios. Such insights can motivate deeper analysis of model architectures and training strategies, ultimately supporting the development of more robust countermeasures against synthetic speech generated by these models.

Overall, this work lays the foundation for inclusive, responsible, and sustainable speech technologies. The key outputs, namely IndicSynth, Task Lens, FAtNet, EcoSpeak, and GreenVoice, provide practical tools, frameworks, and insights to guide future research in this direction. The creation and evaluation of IndicSynth, in particular, have been recognized with an Outstanding Paper Award at ACL 2025, underscoring the relevance and significance of this work to the research community.

# REFERENCES


[1] **Abraham, B.**, **D. Goel**, **D. Siddarth**, **K. Bali**, **M. Chopra**, **M. Choudhury**, **P. Joshi**, **P. Jyoti**, **S. Sitaram**, and **V. Seshadri**, Crowdsourcing speech data for low-resource languages from low-income workers. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.343/.

[2] **Adiga, D.**, **R. Kumar**, **A. Krishna**, **P. Jyothi**, **G. Ramakrishnan**, and **P. Goyal**, Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2021. URL https://aclanthology.org/2021.findings-acl.447/.

[3] **Ahamad, A.**, **A. Anand**, and **P. Bhargava**, AccentDB: A database of non-native English accents to assist neural speech recognition. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.659/.

[4] **AI4Bharat** (2020). Nptel2020 indian english speech dataset. https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset.

[5] **Akram, A.**, **M. Stanojevic**, **M. Ehghaghi**, and **J. Novikova** (2024). Zero-shot multi-lingual speaker verification in clinical trials. *ArXiv*, **abs/2404.01981**. URL https://api.semanticscholar.org/CorpusID:268857288.

[6] **Alharthi, D.**, **R. Sharma**, **H. Dhamyal**, **S. Maiti**, **B. Raj**, and **R. Singh** (2023). Evaluating speech synthesis by training recognizers on synthetic speech. URL https://arxiv.org/abs/2310.00706.

[7] **Auckenthaler, R.**, **M. J. Carey**, and **J. S. D. Mason**, Language dependency in text-independent speaker verification. *In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1. 2001.

[8] **Ba, Z.**, **Q. Wen**, **P. Cheng**, **Y. Wang**, **F. Lin**, **L. Lu**, and **Z. Liu**, Transferring audio deepfake detection capability across languages. *In Proceedings of the ACM Web Conference 2023*, WWW '23. Association for Computing Machinery, New York, NY, USA, 2023. ISBN 9781450394161. URL https://doi.org/10.1145/3543507.3583222.

[9] **Ba, Z.**, **Q. Wen**, **P. Cheng**, **Y. Wang**, **F. Lin**, **L. Lu**, and **Z. Liu**, Transferring audio deepfake detection capability across languages. *In Proceedings of the ACM Web Conference 2023*, WWW '23. Association for Computing Machinery, New York, NY, USA, 2023. ISBN 9781450394161. URL https://doi.org/10.1145/3543507.3583222.

[10] **Babu, A.**, **C. Wang**, **A. Tjandra**, **K. Lakhotia**, **Q. Xu**, **N. Goyal**, **K. Singh**, **P. von Platen**, **Y. Saraf**, **J. Pino**, **A. Baevski**, **A. Conneau**, and **M. Auli**, XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *In Proc. Interspeech 2022*. 2022.

[11] **Bakhturina, E.**, **V. Lavrukhin**, and **B. Ginsburg**, A toolbox for construction and analysis of speech datasets. *In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL `https://openreview.net/forum?id=oJ0oHQtAld`.

[12] **Belinkov, Y.**, **A. Ali**, and **J. Glass**, Analyzing Phonetic and Graph- emic Representations in End-to-End Automatic Speech Recognition. *In Proc. Interspeech 2019*. 2019.

[13] **Besacier, L.**, **E. Barnard**, **A. Karpov**, and **T. Schultz** (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, **56**, 85–100.

[14] **Bhanushali, A.**, **G. Bridgman**, **D. G**, **P. Ghosh**, **P. Kumar**, **S. Kumar**, **A. Raj Kolladath**, **N. Ravi**, **A. Seth**, **A. Seth**, **A. Singh**, **V. Sukhadia**, **U. S**, **S. Udupa**, and **L. V. S. V. D. Prasad**, Gram vaani asr challenge on spontaneous telephone speech recordings in regional variations of hindi. *In Interspeech 2022*. 2022. ISSN 2958-1796.

[15] **Bird, J. J.** and **A. Lotfi** (2023). Real-time detection of ai-generated speech for deepfake voice conversion. *ArXiv*, **abs/2308.12734**. URL `https://api.semanticscholar.org/CorpusID:261101196`.

[16] **BizAugmentor Global Services** and **GIZ** (2020). Study on open voice data in indian languages. `https://www.bmz-digital.global/wp-content/uploads/2022/08/Study-on-Open-Voice-Data-in-Indian-Languages_GIZ-BizAugmentor.pdf`.

[17] **Brignatz, V.**, **J. Duret**, **D. Matrouf**, and **M. Rouvier**, Language adaptation for speaker recognition systems using contrastive learning. *In* **A. Karpov** and **R. Potapova** (eds.), *Speech and Computer*. Springer International Publishing, Cham, 2021.

[18] **Bu, H.**, **J. Du**, **X. Na**, **B. Wu**, and **H. Zheng** (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline.

[19] **Busso, C.**, **M. Bulut**, **C.-C. Lee**, **A. Kazemzadeh**, **E. Mower**, **S. Kim**, **J. N. Chang**, **S. Lee**, and **S. S. Narayanan** (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **42**(4), 335–359.

[20] **Casanova, E.**, **J. Weber**, **C. D. Shulby**, **A. C. Junior**, **E. Gölge**, and **M. A. Ponti**, Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. *In International Conference on Machine Learning*. PMLR, 2022.

[21] **Casanova, E.** *et al.* (2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. *arXiv preprint arXiv:2406.04904*.

[22] **Chen, J.**, **C. Chu**, **S. Li**, and **T. Kawahara**, Data selection using spoken language identification for low-resource and zero-resource speech recognition. *In 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. APSIPA, Tokyo, Japan, 2024.

[23] **Chen, M.**, **P.-A. Duquenne**, **P. Andrews**, **J. Kao**, **A. Mourachko**, **H. Schwenk**, and **M. R. Costa-jussà**, BLASER: A text-free speech-to-speech translation evaluation metric. *In* **A. Rogers**, **J. Boyd-Graber**, and **N. Okazaki** (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 2023. URL https://aclanthology.org/2023.acl-long.504/.

[24] **Chen, Z.**, **S. Wang**, and **Y. Qian**, Adversarial Domain Adaptation for Speaker Verification Using Partially Shared Network. *In Proc. Interspeech*. 2020. URL http://dx.doi.org/10.21437/Interspeech.2020-2226.

[25] **Chen, Z.**, **S. Wang**, and **Y. Qian**, Self-supervised learning based domain adaptation for robust speaker verification. *In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

[26] **Choe, J.**, **Y. Chen**, **M. P. Y. Chan**, **A. Li**, **X. Gao**, and **N. Holliday**, Language-specific effects on automatic speech recognition errors for world englishes. *In* **N. Calzolari**, **C.-R. Huang**, **H. Kim**, **J. Pustejovsky**, **L. Wanner**, **K.-S. Choi**, **P.-M. Ryu**, **H.-H. Chen**, **L. Donatelli**, **H. Ji**, **S. Kurohashi**, **P. Paggio**, **N. Xue**, **S. Kim**, **Y. Hahm**, **Z. He**, **T. K. Lee**, **E. Santus**, **F. Bond**, and **S.-H. Na** (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022. URL https://aclanthology.org/2022.coling-1.628.

[27] **Chojnacka, R.**, **J. Pelecanos**, **Q. Wang**, and **I. Lopez-Moreno** (2021). Speakerstew: Scaling to many languages with a triaged multilingual text-dependent and text-independent speaker verification system. *CoRR*, **abs/2104.02125**.

[28] **Chojnacka, R.**, **J. Pelecanos**, **Q. Wang**, and **I. L. Moreno**, SpeakerStew: Scaling to Many Languages with a Triaged Multilingual Text-Dependent and Text-Independent Speaker Verification System. *In Proc. Interspeech 2021*. 2021.

[29] **Christop, I.**, **T. Kuczyński**, and **M. Kubis**, Cloneval: An open voice cloning benchmark. 2025. URL https://arxiv.org/abs/2504.20581.

[30] **Chrupała, G.**, **B. Higy**, and **A. Alishahi**, Analyzing analytical methods: The case of phonology in neural models of spoken language. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020.

[31] **Chung, J. S.**, **A. Nagrani**, and **A. Zisserman**, Voxceleb2: Deep speaker recognition. *In INTERSPEECH*. 2018.

[32] **Ciobanu, I.-P.**, **A.-I. Hiji**, **N.-C. Ristea**, **P. Irofti**, **C. Rusu**, and **R. T. Ionescu** (2025). Xmad-bench: Cross-domain multilingual audio deepfake benchmark.

[33] **Computing, S. M.** (2020). Malayalam speech corpus. https://blog.smc.org.in/malayalam-speech-corpus/. Accessed: 2025-05-23.

[34] **Conneau, A.**, **M. Ma**, **S. Khanuja**, **Y. Zhang**, **V. Axelrod**, **S. R. Dalmia**, **J. Riesa**, **C. Rivera**, and **A. Bapna** (2022). Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*. URL https://arxiv.org/abs/2205.12446.

[35] **Cooper, E.**, **W.-C. Huang**, **T. Toda**, and **J. Yamagishi**, Generalization ability of mos prediction networks. *In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022.

[36] **Courty, B.**, **V. Schmidt**, **S. Luccioni**, **Goyal-Kamal**, **MarionCoutarel**, **B. Feld**, **J. Lecourt**, **LiamConnell**, **A. Saboni**, **Inimaz**, **supatomic**, **M. Léval**, **L. Blanche**, **A. Cruveiller**, **ouminasara**, **F. Zhao**, **A. Joshi**, **A. Bogroff**, **H. de Lavoreille**, **N. Laskaris**, **E. Abati**, **D. Blank**, **Z. Wang**, **A. Catovic**, **M. Alencon**, **Michał Stęchły**, **C. Bauer**, **L. O. N. de Araújo**, **JPW**, and **MinervaBooks** (2024). mlco2/codecarbon: v2.4.1. URL https://doi.org/10.5281/zenodo.11171501.

[37] **Demirsahin, I.**, **O. Kjartansson**, **A. Gutkin**, and **C. Rivera**, Open-source Multi-speaker Corpora of the English Accents in the British Isles. *In* **N. Calzolari**, **F. Béchet**, **P. Blache**, **K. Choukri**, **C. Cieri**, **T. Declerck**, **S. Goggi**, **H. Isahara**, **B. Maegaard**, **J. Mariani**, **H. Mazo**, **A. Moreno**, **J. Odijk**, and **S. Piperidis** (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.804.

[38] **Desplanques, B.**, **J. Thienpondt**, and **K. Demuynck**, ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *In* **H. Meng**, **B. Xu**, and **T. F. Zheng** (eds.), *Interspeech 2020*. ISCA, 2020.

[39] **Desplanques, B.** *et al.* (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

[40] **Diwan, A.**, **R. Vaideeswaran**, **S. Shah**, **A. Singh**, **S. Raghavan**, **S. Khare**, **V. Unni**, **S. Vyas**, **A. Rajpuria**, **C. Yarra**, **A. Mittal**, **P. K. Ghosh**, **P. Jyothi**, **K. Bali**, **V. Seshadri**, **S. Sitaram**, **S. Bharadwaj**, **J. Nanavati**, **R. Nanavati**, and **K. Sankaranarayanan**, Mucs 2021: Multilingual and code-switching asr challenges for low resource indian languages. *In Proceedings of Interspeech 2021*. 2021.

[41] **Diwan, A.**, **R. Vaideeswaran**, **S. Shah**, **A. Singh**, **S. Raghavan**, **S. Khare**, **V. Unni**, **S. Vyas**, **A. Rajpuria**, **C. Yarra**, **A. Mittal**, **P. K. Ghosh**, **P. Jyothi**, **K. Bali**, **V. Seshadri**, **S. Sitaram**, **S. Bharadwaj**, **J. Nanavati**, **R. Nanavati**, and **K. Sankaranarayanan**, MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. *In Proc. Interspeech 2021*. 2021.

[42] **Douwes, C.**, **P. Esling**, and **J.-P. Briot** (2021). Energy consumption of deep generative audio models. URL https://arxiv.org/abs/2107.02621.

[43] **ELRA** (2018). GlobalPhone 2000 speaker package corpus. ELRA Catalogue, ID: ELRA-S0400. ISLRN: 331-592-378-424-7; available since 2018-02-10 (submission date: 2018-10-09) :contentReference[oaicite:0]index=0. URL: https://catalogue.elra.info/en-us/repository/browse/ELRA-S0400/ (accessed 2025-05-30).

[44] **ELRA** (2024). Urdu Speech Recognition Corpus (Desktop). ELRA Catalogue, Corpus No. ELRA-S0228-117. ISLRN: 739-446-795-223-8. Available: https://catalogue.elra.info/en-us/repository/browse/ELRA-S0228_117/. Accessed: 30 May 2025.

[45] **Eren, G.** and **The Coqui TTS Team** (2021). Coqui TTS. URL https://github.com/coqui-ai/TTS.

[46] **Estevez, M.** and **L. Ferrer**, Study on the fairness of speaker verification systems across accent and gender groups. *In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023.

[47] **Frank, J. C.** and **L. Schönherr** (2021). Wavefake: A data set to facilitate audio deepfake detection. *ArXiv*, **abs/2111.02813**. URL https://api.semanticscholar.org/CorpusID:242757450.

[48] **Gao, Z.**, **Y. Song**, **I. McLoughlin**, **W. Guo**, and **L. Dai**, An improved deep embedding learning method for short duration speaker verification. *In Proc. Interspeech 2018*. 2018.

[49] **Gebru, T.**, **J. Morgenstern**, **B. Vecchione**, **J. W. Vaughan**, **H. Wallach**, **H. D. III**, and **K. Crawford** (2021). Datasheets for datasets. URL https://arxiv.org/abs/1803.09010.

[50] **Goel, A.**, **M. Hira**, and **A. Gupta**, Exploring multilingual unseen speaker emotion recognition: Leveraging co-attention cues in multitask learning. *In Interspeech 2024*. 2024. ISSN 2958-1796.

[51] **Gupta, M.**, **M. Dutta**, and **C. K. Maurya** (2025). Benchmarking hindi-to-english direct speech-to-speech translation with synthetic data. *Language Resources and Evaluation*, **59**, 2613–2651. URL https://doi.org/10.1007/s10579-025-09827-2.

[52] **Guzewich, P.**, **S. Zahorian**, **X. Chen**, and **H. Zhang**, Cross-corpora convolutional deep neural network dereverberation preprocessing for speaker verification and speech enhancement. *In Proc. Interspeech 2018*. 2018.

[53] **Hansen, J. H. L.** and **T. Hasan** (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, **32**(6), 74–99.

[54] **Haut, K.**, **C. Wohn**, **V. Antony**, **A. Goldfarb**, **M. Welsh**, **D. Sumanthiran**, **M. Rafayet Ali**, and **E. Hoque**, Demographic feature isolation for bias research using deepfakes. *In Proceedings of the 30th ACM International Conference on Multimedia*, MM '22. Association for Computing Machinery, New York, NY, USA, 2022. ISBN 9781450392037. URL https://doi.org/10.1145/3503161.3549204.

[55] **Havard, W. N.**, **J.-P. Chevrot**, and **L. Besacier**, Word recognition, competition, and activation in a model of visually grounded speech. *In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 2019.

[56] **He, F.**, **S.-H. C. Chu**, **O. Kjartansson**, **C. Rivera**, **A. Katanova**, **A. Gutkin**, **I. Demirsahin**, **C. Johny**, **M. Jansche**, **S. Sarin**, and **K. Pipatsrisawat**, Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.800/.

[57] **He, F.**, **S.-H. C. Chu**, **O. Kjartansson**, **C. Rivera**, **A. Katanova**, **A. Gutkin**, **I. Demirsahin**, **C. Johny**, **M. Jansche**, **S. Sarin**, and **K. Pipatsrisawat**, Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. *In* **N. Calzolari**, **F. Béchet**, **P. Blache**, **K. Choukri**, **C. Cieri**, **T. Declerck**, **S. Goggi**, **H. Isahara**, **B. Maegaard**, **J. Mariani**, **H. Mazo**, **A. Moreno**, **J. Odijk**, and **S. Piperidis** (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.800.

[58] **He, K.**, **X. Zhang**, **S. Ren**, and **J. Sun**, Deep residual learning for image recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[59] **Huang, P.-Y.**, **M. Patrick**, **J. Hu**, **G. Neubig**, **F. Metze**, and **A. Hauptmann**, Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2021.

[60] **Huang, W.-C.**, **E. Cooper**, **J. Yamagishi**, and **T. Toda**, Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech. *In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022.

[61] **IIT-M** (2021). Hindi–tamil–english asr challenge. https://sites.google.com/view/indian-language-asrchallenge/home. Accessed: 2025-05-30.

[62] **IIT Madras** (2025). IndicTTS: Text-to-speech corpus for indian languages. https://www.iitm.ac.in/donlab/indictts. Accessed: 2025-05-30.

[63] **Jain, S.**, **A. Sankar**, **D. Choudhary**, **D. Suman**, **N. Narasimhan**, **M. S. U. R. Khan**, **A. Kunchukuttan**, **M. M. Khapra**, and **R. Dabre** (2024). Bhasaanuvaad: A speech translation dataset for 13 indian languages. *arXiv preprint arXiv:2411.04699*. URL https://arxiv.org/abs/2411.04699.

[64] **Javed, T.**, **K. Bhogale**, **A. Raman**, **P. Kumar**, **A. Kunchukuttan**, and **M. Khapra** (2023). Indicsuperb: A speech processing universal performance benchmark for indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 12942–12950.

[65] **Javed, T.**, **K. Bhogale**, **A. Raman**, **P. Kumar**, **A. Kunchukuttan**, and **M. M. Khapra** (2023). Indicsuperb: A speech processing universal performance benchmark for indian languages.

[66] **Javed, T.**, **S. Joshi**, **V. Nagarajan**, **S. Sundaresan**, **J. Nawale**, **A. Raman**, **K. Bhogale**, **P. Kumar**, and **M. M. Khapra**, Svarah: Evaluating english asr systems on indian accents. *In Interspeech 2023*. 2023. ISSN 2958-1796.

[67] **Javed, T.**, **J. Nawale**, **S. Joshi**, **E. George**, **K. Bhogale**, **D. Mehendale**, and **M. M. Khapra**, Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems. *In Interspeech 2024*. 2024. ISSN 2958-1796.

[68] **Jia, Y.** *et al.* (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, **31**.

[69] **Joshi, P.**, **S. Santy**, **A. Budhiraja**, **K. Bali**, and **M. Choudhury**, The state and fate of linguistic diversity and inclusion in the NLP world. *In* **D. Jurafsky**, **J. Chai**, **N. Schluter**, and **J. Tetreault** (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020. URL `https://aclanthology.org/2020.acl-main.560/`.

[70] **Ju, Y.**, **S. Hu**, **S. Jia**, **G. H. Chen**, and **S. Lyu**, Improving Fairness in Deepfake Detection . *In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Los Alamitos, CA, USA, 2024. URL `https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00459`.

[71] **Jung, J.-W.**, **H.-S. Heo**, **J.-H. Kim**, **H.-J. Shim**, and **H.-J. Yu**, Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. 2019.

[72] **Jung, J.-w.**, **H.-S. Heo**, **H. Tak**, **H.-j. Shim**, **J. S. Chung**, **B.-J. Lee**, **H.-J. Yu**, and **N. 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)*. 2022.

[73] **Jung, J.-w.**, **S.-b. Kim**, **H.-j. Shim**, **J.-h. Kim**, and **H.-J. Yu** (2020). Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Proc. Interspeech 2020*, 3583–3587.

[74] **Jung, J.-W.**, **Y. Kim**, **H.-S. Heo**, **B.-J. Lee**, **Y. Kwon**, and **J. S. Chung**, Pushing the limits of raw waveform speaker recognition. 2022.

[75] **Kalluri, S. B.**, **D. Vijayasenan**, **S. Ganapathy**, **R. R. M**, and **P. Krishnan**, Nisp: A multi-lingual multi-accent dataset for speaker profiling. *In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

[76] **Kalluri, S. B.**, **D. Vijayasenan**, **S. Ganapathy**, **R. R. M**, and **P. Krishnan**, Nisp: A multi-lingual multi-accent dataset for speaker profiling. *In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

[77] **Kalluri, S. B.** *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)*. IEEE, 2021.

[78] **Kawa, P.**, **M. Plata**, and **P. Syga**, Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. 2022.

[79] **Khalid, H.**, **S. Tariq**, and **S. S. Woo** (2021). Fakeavceleb: A novel audio-video multimodal deepfake dataset. *ArXiv*, **abs/2108.05080**. URL https://api.semanticscholar.org/CorpusID:236976127.

[80] **Khoury, E.**, **L. Shafey**, and **S. Marcel**, Spear: An open source toolbox for speaker recognition based on bob. 2014. ISBN 978-1-4799-2893-4.

[81] **Kjartansson, O.**, **S. Sarin**, **K. Pipatsrisawat**, **M. Jansche**, and **L. Ha**, Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. *In Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*. Gurugram, India, 2018. URL http://dx.doi.org/10.21437/SLTU.2018-11.

[82] **Konduri, S.**, **K. V. Pendyala**, and **V. S. Pendyala** (2024). Kritisamhita: A machine learning dataset of south indian classical music audio clips with tonic classification. *Data in Brief*, **55**, 110730. URL https://doi.org/10.1016/j.dib.2024.110730.

[83] **Korshunov, P.** and **S. 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.

[84] **Korshunov, P.** and **S. 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.

[85] **Larasati, R.** (2025). Inclusivity of ai speech in healthcare: A decade look back.

[86] **Lee, K. A.**, **K. Okabe**, **H. Yamamoto**, **Q. Wang**, **L. Guo**, **T. Koshinaka**, **J. Zhang**, **K. Ishikawa**, and **K. Shinoda**, NEC-TT Speaker Verification System for SRE'19 CTS Challenge. *In Proc. Interspeech 2020*. 2020.

[87] **Leng, Y.**, **X. Tan**, **S. Zhao**, **F. K. Soong**, **X.-Y. Li**, and **T. Qin** (2021). Mbnet: Mos prediction for synthesized speech with mean-bias network. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 391–395. URL https://api.semanticscholar.org/CorpusID:232076409.

[88] **Lesenfants, D.**, **J. Vanthornhout**, **E. Verschueren**, **L. Decruy**, and **T. Francart** (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *Hearing Research*, **380**.

[89] **Li, J.**, **W. Liu**, and **T. Lee**, EDITnet: A Lightweight Network for Unsupervised Domain Adaptation in Speaker Verification. *In Proc. Interspeech 2022*. 2022.

[90] **Li, J.**, **W. Tu**, and **L. 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)*. IEEE, 2023.

[91] **Li, K.**, **C. Shen**, **Y. Liu**, **J. Han**, **K. Zheng**, **X. Zou**, **Z. Wang**, **S. Zhang**, **X. Du**, **H. Luo**, **Y. Jin**, **X. Xing**, **Z. Ma**, **Y. Liu**, **Y. Zhang**, **J. Fang**, **K. Wang**, **Y. Yan**, **G. Deng**, **H. Li**, **Y. Li**, **X. Zhuang**, **T. Chen**, **Q. Wen**, **T. Zhang**, **Y. Liu**, **H. Hu**, **Z. Wu**, **X. Hu**, **E.-S. Chng**, **W. Xu**, **X. Wang**, **W. Dong**, and **X. Li** (2025). Audiotrust: Benchmarking the multifaceted trustworthiness of audio large language models. URL `https://arxiv.org/abs/2505.16211`.

[92] **Li, N.**, **D. Tuo**, **D. Su**, **Z. Li**, and **D. Yu**, Deep discriminative embeddings for duration robust speaker verification. *In Proc. Interspeech 2018*. 2018.

[93] **Liu, S.** (2024). Zero-shot Voice Conversion with Diffusion Transformers. *arXiv preprint arXiv:2411.09943*.

[94] **Liu, X.**, **X. Wang**, **M. Sahidullah**, **J. Patino**, **H. Delgado**, **T. Kinnunen**, **M. Todisco**, **J. Yamagishi**, **N. Evans**, **A. Nautsch**, and **K. A. 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.

[95] **Lu, L.**, **Y. Dong**, **X. Zhao**, **J. Liu**, and **H. Wang**, The effect of language factors for robust speaker recognition. *In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009.

[96] **Lyons, J.** (). Mel frequency cepstral coefficient (mfcc) tutorial.

[97] **Maiti, S.**, **Y. Peng**, **T. Saeki**, and **S. Watanabe** (2022). Speechlmscore: Evaluating speech generation using speech language model. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. URL `https://api.semanticscholar.org/CorpusID:254535560`.

[98] **Malek, A.** (2020). Signal framing.

[99] **Mandalapu, H.**, **T. M. Elbo**, **R. Ramachandra**, and **C. Busch**, Cross-lingual speaker verification: Evaluation on x-vector method. *In* **S. Yildirim Yayilgan**, **I. S. Bajwa**, and **F. Sanfilippo** (eds.), *Intelligent Technologies and Applications*. Springer International Publishing, Cham, 2021.

[100] **Mazumder, M.**, **C. Banbury**, **X. Yao**, **B. Karlaş**, **W. G. Rojas**, **S. Diamos**, **G. Diamos**, **L. He**, **A. Parrish**, **H. R. Kirk**, **J. Quaye**, **C. Rastogi**, **D. Kiela**, **D. Jurado**, **D. Kanter**, **R. Mosquera**, **J. Ciro**, **L. Aroyo**, **B. Acun**, **L. Chen**, **M. S. Raje**, **M. Bartolo**, **S. Eyuboglu**, **A. Ghorbani**, **E. Goodman**, **O. Inel**, **T. Kane**, **C. R. Kirkpatrick**, **T.-S. Kuo**, **J. Mueller**, **T. Thrush**, **J. Vanschoren**, **M. Warren**, **A. Williams**, **S. Yeung**, **N. Ardalani**, **P. Paritosh**, **C. Zhang**, **J. Zou**, **C.-J. Wu**, **C. Coleman**, **A. Ng**, **P. Mattson**, and **V. J. Reddi**, Dataperf: benchmarks for data-centric ai development. *In Proceedings of the 37th*

*International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., Red Hook, NY, USA, 2023.

[101] **McFee, B.**, **C. Raffel**, **D. Liang**, **D. Ellis**, **M. Mcvicar**, **E. Battenberg**, and **O. Nieto**, librosa: Audio and music signal analysis in python. 2015.

[102] **McFee, B.**, **C. Raffel**, **D. Liang**, **D. P. W. Ellis**, **M. McVicar**, **E. Battenberg**, and **O. Nieto**, librosa: Audio and music signal analysis in python. 2015.

[103] **Metavoiceio** (2024). Metavoice Source Code. `https://github.com/metavoiceio/metavoice-src`. Accessed: 2024-09-12.

[104] **Microsoft** (). Microsoft Speech Corpus (Indian languages). `https://www.microsoft.com/en-us/download/details.aspx?id=105292`. Version 1.0; published July 15, 2024; Online; accessed May 23, 2025.

[105] **Microsoft** (2023). Microsoft speech corpus (indian languages). https://www.microsoft.com/en-za/download/details.aspx?id=105292.

[106] **Microsoft** (2024). Rajasthani Hindi Speech Data. `https://www.microsoft.com/en-gb/download/details.aspx?id=105385`. Published 2024; accessed May 23, 2025.

[107] **Moshnoi, I.** (). All you need is attention — computer vision edition. 04/19/2018.

[108] **Munir, S.**, **W. Sajjad**, **M. Raza**, **E. Abbas**, **A. H. Azeemi**, **I. A. Qazi**, and **A. A. Raza**, Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset. *In* **L.-W. Ku**, **A. Martins**, and **V. Srikumar** (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 2024. URL `https://aclanthology.org/2024.findings-acl.861/`.

[109] **Müller, N.**, **P. Czempin**, **F. Diekmann**, **A. Froghyar**, and **K. Böttinger**, Does Audio Deepfake Detection Generalize? *In Interspeech 2022*. 2022. ISSN 2958-1796.

[110] **Müller, N.**, **N. Evans**, **H. Tak**, **P. Sperl**, and **K. Böttinger**, Harder or different? understanding generalization of audio deepfake detection. 2024.

[111] **Müller, N. M.**, **N. Evans**, **H. Tak**, **P. Sperl**, and **K. Böttinger**, Harder or Different? Understanding Generalization of Audio Deepfake Detection. *In Interspeech 2024*. 2024. ISSN 2958-1796.

[112] **Müller, N. M.**, **P. Kawa**, **W. H. Choong**, **E. Casanova**, **E. Gölge**, **T. Müller**, **P. Syga**, **P. Sperl**, and **K. Böttinger**, Mlaad: The multi-language audio anti-spoofing dataset. *In 2024 International Joint Conference on Neural Networks (IJCNN)*. 2024.

[113] **Nagrani, A.**, **J. S. Chung**, **W. Xie**, and **A. Zisserman** (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, **60**, 101027. ISSN 0885-2308.

[114] **Nagrani, A.**, **J. S. Chung**, and **A. Zisserman**, Voxceleb: a large-scale speaker identification dataset. *In INTERSPEECH*. 2017.

[115] **Nahid, M. M. H.**, **M. A. Islam**, and **M. S. Islam** (2018). Bengali speech recognition - bangla real number audio dataset. Mendeley Data, V6.

[116] **Nahum, M.**, **I. Nelken**, and **M. Ahissar** (2008). Low-level information and high-level perception: The case of speech in noise. *PLoS biology*, **6**, e126.

[117] **Nexdata AI** (). 759 hours – hindi(india) scripted monologue smartphone speech dataset. https://www.nexdata.ai/datasets/speechrecog/946. Commercial license; accessed 2025-05-30.

[118] **Nguyen, B.**, **S. Shi**, **R. Ofman**, and **T. Le**, What you read isn't what you hear: Linguistic sensitivity in deepfake speech detection. *In* **C. Christodoulopoulos**, **T. Chakraborty**, **C. Rose**, and **V. Peng** (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 2025. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.794/.

[119] **Okabe, K.**, **T. Koshinaka**, and **K. Shinoda**, Attentive statistics pooling for deep speaker embedding. *In Proc. Interspeech 2018*. 2018.

[120] **Oleg, K.**, **S. Novoselov**, **T. Pekhovsky**, **K. Simonchik**, and **G. Lavrentyeva**, Usage of dnn in speaker recognition: Advantages and problems. *In Advances in Neural Networks – ISNN 2016*, volume 9719. 2016. ISBN 978-3-319-40662-6.

[121] **OpenSLR** (). SLR122: Kashmiri Data Corpus. https://www.openslr.org/122/. Online; accessed May 23, 2025.

[122] **Panayotov, V.**, **G. Chen**, **D. Povey**, and **S. Khudanpur**, Librispeech: An asr corpus based on public domain audio books. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015.

[123] **Panayotov, V.**, **G. Chen**, **D. Povey**, and **S. Khudanpur**, Librispeech: An asr corpus based on public domain audio books. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015.

[124] **Passoni, R.**, **F. Ronchini**, **L. Comanducci**, **R. Serizel**, and **F. Antonacci** (2025). Diffused responsibility: Analyzing the energy consumption of generative text-to-audio diffusion models. URL https://arxiv.org/abs/2505.07615.

[125] **Paszke, A.**, **S. Gross**, **S. Chintala**, **G. Chanan**, **E. Yang**, **Z. DeVito**, **Z. Lin**, **A. Desmaison**, **L. Antiga**, and **A. Lerer**, Automatic differentiation in pytorch. *In NIPS-W*. 2017.

[126] **Paszke, A.**, **S. Gross**, **F. Massa**, **A. Lerer**, **J. Bradbury**, **G. Chanan**, **T. Killeen**, **Z. Lin**, **N. Gimelshein**, **L. Antiga**, **A. Desmaison**, **A. Köpf**, **E. Yang**, **Z. DeVito**, **M. Raison**, **A. Tejani**, **S. Chilamkurthy**, **B. Steiner**, **L. Fang**, **J. Bai**, and **S. Chintala**, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[127] **Popov, V.**, **I. Vovk**, **V. Gogoryan**, **T. Sadekova**, **M. S. Kudinov**, and **J. 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.

[128] **Pukhraj P. Shrishrimal, V. B. W., Ratnadeep R. Deshmukh** (2012). Indian language speech database: A review. *International Journal of Computer Applications*, **47**(5), 17–21. ISSN 0975-8887. URL https://ijcaonline.org/archives/volume47/number5/7184-9893/.

[129] **Purohit, R. M.**, **A. J. Shah**, and **H. A. Patil**, Ggmddc: An audio deepfake detection multilingual dataset. *In 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2024.

[130] **Qi, X.**, **H. Gu**, **J. Yi**, **J. Tao**, **Y. Ren**, **J. He**, and **S. Zeng**, Madd: A multi-lingual multi-speaker audio deepfake detection dataset. *In 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2024.

[131] **Qin, X.**, **C. Wang**, **Y. Ma**, **M. Liu**, **S. Zhang**, and **M. Li**, Our Learned Lessons from Cross-Lingual Speaker Verification: The CRMI-DKU System Description for the Short-Duration Speaker Verification Challenge 2021. *In Proc. Interspeech 2021*. 2021.

[132] **Qin, Z.** *et al.* (2023). Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

[133] **Rabhi, M.**, **S. Bakiras**, and **R. Di Pietro** (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, **250**, 123941. ISSN 0957-4174. URL https://www.sciencedirect.com/science/article/pii/S0957417424008078.

[134] **Ravanelli, M.**, **T. Parcollet**, **P. Plantinga**, **A. Rouhe**, **S. Cornell**, **L. Lugosch**, **C. Subakan**, **N. Dawalatabad**, **A. Heba**, **J. Zhong**, **J.-C. Chou**, **S.-L. Yeh**, **S.-W. Fu**, **C.-F. Liao**, **E. Rastorgueva**, **F. Grondin**, **W. Aris**, **H. Na**, **Y. Gao**, **R. D. Mori**, and **Y. Bengio** (2021). SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

[135] **Ravanelli, M.** *et al.* (2021). Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

[136] **Reddy, C. K. A.**, **V. Gopal**, and **R. 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)*. IEEE, Singapore, 2022.

[137] **Rohdin, J.**, **T. Stafylakis**, **A. Silnova**, **H. Zeinali**, **L. Burget**, and **O. Plchot**, Speaker verification using end-to-end adversarial language adaptation. *In ICASSP 2019*.

[138] **Rohdin, J.**, **T. Stafylakis**, **A. Silnova**, **H. Zeinali**, **L. Burget**, and **O. Plchot**, Speaker verification using end-to-end adversarial language adaptation. *In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[139] **Saha, S.**, **M. Sahidullah**, and **S. Das**, Exploring green ai for audio deepfake detection. *In 2024 32nd European Signal Processing Conference (EUSIPCO)*. 2024.

[140] **Salesky, E.**, **J. Mäder**, and **S. Klinger**, Assessing evaluation metrics for speech-to-speech translation. *In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2021.

[141] **Sankar, A.**, **S. Anand**, **P. S. Varadhan**, **S. Thomas**, **M. Singal**, **S. Kumar**, **D. Mehendale**, **A. Krishana**, **G. Raju**, and **M. Khapra** (2024). Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. URL https://arxiv.org/abs/2409.05356.

[142] **Schwartz, R.**, **J. Dodge**, **N. A. Smith**, and **O. Etzioni** (2020). Green ai. *Commun. ACM*, **63**(12), 54–63. ISSN 0001-0782. URL https://doi.org/10.1145/3381831.

[143] **SHAH, A. J.**, **R. M. Purohit**, **D. H. Vaghera**, and **H. Patil**, MLADDC: Multi-lingual audio deepfake detection corpus. *In Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*. 2024. URL https://openreview.net/forum?id=ic3HvoOTeU.

[144] **Shain, C.** and **M. Elsner**, Acquiring language from speech by learning to remember and predict. *In Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 2020. URL https://www.aclweb.org/anthology/2020.conll-1.15.

[145] **Sharma, D.**, EcoSpeak: Cost-efficient bias mitigation for partially cross-lingual speaker verification. *In* **K. Duh**, **H. Gomez**, and **S. Bethard** (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 2024. URL https://aclanthology.org/2024.findings-naacl.27/.

[146] **Sharma, D.** and **A. B. Buduru**, FAtNet: Cost-effective approach towards mitigating the linguistic bias in speaker verification systems. *In* **M. Carpuat**, **M.-C. de Marneffe**, and **I. V. Meza Ruiz** (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 2022. URL https://aclanthology.org/2022.findings-naacl.93/.

[147] **Sharma, D.** and **A. B. Buduru**, FAtNet: Cost-effective approach towards mitigating the linguistic bias in speaker verification systems. *In* **M. Carpuat**, **M.-C. de Marneffe**, and **I. V. Meza Ruiz** (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 2022. URL https://aclanthology.org/2022.findings-naacl.93/.

[148] **Shi, X.**, **F. Yu**, **Y. Lu**, **Y. Liang**, **Q. Feng**, **D. Wang**, **Y. Qian**, and **L. Xie**, The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[149] **Shilin, I.**, **L. Kovriguina**, **D. Mouromtsev**, **G. Wohlgenannt**, and **R. Ivanitskiy**, A method for dataset creation for dialogue state classification in voice control systems for the internet of things. 2018.

[150] **Singh, Y.** and **A. Biswas** (2021). Indian regional music dataset. URL https://doi.org/10.5281/zenodo.6546501.

[151] **Singh, Y.**, **L. Waikhom**, **V. Meena**, and **A. Biswas** (2022). Indian folk music dataset. URL https://zenodo.org/records/6584020.

[152] **Snyder, D.**, **D. Garcia-Romero**, **A. McCree**, **G. Sell**, **D. Povey**, and **S. Khudanpur**, Spoken language recognition using x-vectors. *In The Speaker and Language Recognition Workshop (Odyssey 2018)*. 2018.

[153] **Snyder, D.**, **D. Garcia-Romero**, **G. Sell**, **D. Povey**, and **S. Khudanpur**, X-vectors: Robust dnn embeddings for speaker recognition. 2018.

[154] **Sodimana, K.**, **P. Silva**, **S. Sarin**, **O. Kjartansson**, **M. Jansche**, **K. Pipatsrisawat**, and **L. Ha** (2018). A step-by-step process for building tts voices using open source data and frameworks for bangla, javanese, khmer, nepali, sinhala, and sundanese.

[155] **Srinivasa Varadhan, P.**, **A. Sankar**, **G. Raju**, and **M. M. Khapra**, Rasa: Building expressive speech synthesis systems for indian languages in low-resource settings. *In Interspeech 2024*. 2024. ISSN 2958-1796.

[156] **Srivastava, N.**, **R. Mukhopadhyay**, **P. K R**, and **C. V. Jawahar**, IndicSpeech: Text-to-speech corpus for Indian languages. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.789/.

[157] **Staněk, V.**, **K. Srna**, **A. Firc**, and **K. Malinka** (2025). Scdf: A speaker characteristics deepfake speech dataset for bias analysis. URL https://arxiv.org/abs/2508.07944.

[158] **Strubell, E.**, **A. Ganesh**, and **A. McCallum**, Energy and policy considerations for deep learning in NLP. *In* **A. Korhonen**, **D. Traum**, and **L. Màrquez** (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2019. URL https://aclanthology.org/P19-1355/.

[159] **Swayamdipta, S.**, **R. Schwartz**, **N. Lourie**, **Y. Wang**, **H. Hajishirzi**, **N. A. Smith**, and **Y. Choi**, Dataset cartography: Mapping and diagnosing datasets with training dynamics. *In* **B. Webber**, **T. Cohn**, **Y. He**, and **Y. Liu** (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2020. URL https://aclanthology.org/2020.emnlp-main.746.

[160] **T, S.** and **G. T** (2024). Deepfake technology in social media: Social and legal implications in india. *IJFMR*, **6**(6).

[161] **Tak, H.**, **J. Patino**, **M. Todisco**, **A. Nautsch**, **N. Evans**, and **A. Larcher**, End-to-end anti-spoofing with rawnet2. *In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

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

[163] **Tamim, S. M.**, **P. Manjul**, **S. Fernandes**, **N. S.**, **R. M. R.**, **N. K. Choudhary**, and **S. Mohan** (2025). Assamese text to speech corpus. Technical report, Central Institute of Indian Languages, Mysore.

[164] **Thelisson, E.**, **G. Mika**, **Q. Schneiter**, **K. Padh**, and **H. Verma** (2023). Toward responsible ai use: Considerations for sustainability impact assessment. URL https://arxiv.org/abs/2312.11996.

[165] **Thienpondt, J.**, **B. Desplanques**, and **K. Demuynck**, Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization. *In Proc. Interspeech 2020*. 2020.

[166] **Tu, Y.**, **M.-W. Mak**, and **J.-T. Chien**, Variational Domain Adversarial Learning for Speaker Verification. *In Proc. Interspeech 2019*. 2019.

[167] **Udupa, S.**, **J. Bandekar**, **A. Singh**, **D. G**, **S. Kumar**, **S. Badiger**, **A. Nagireddi**, **R. R**, **P. K. Ghosh**, **H. A. Murthy**, **P. Kumar**, **K. Tokuda**, **M. Hasegawa-Johnson**, and **P. Olbrich** (2025). Limmits'24: Multi-speaker, multi-lingual indic tts with voice cloning. *IEEE Open Journal of Signal Processing*, **6**, 293–302.

[168] **Ujjwal**, **H. Garg**, and **M. Joshi** (2020). GACMIS: Genre Automated Classification using Machine Learning of Indian Songs. GitHub repository. https://github.com/ujjwalll/GACMIS (accessed 2025-05-30).

[169] **Ulgen, I. R.**, **Z. Du**, **J. Lu**, **P. Koehn**, and **B. Sisman** (2026). Objective evaluation of prosody and intelligibility in speech synthesis via conditional prediction of discrete tokens. *IEEE Open Journal of Signal Processing*, **7**, 247–256. ISSN 2644-1322. URL http://dx.doi.org/10.1109/OJSP.2026.3653666.

[170] **Valk, J.** and **T. Alumäe**, VoxLingua107: a dataset for spoken language recognition. *In Proc. IEEE SLT Workshop*. 2021.

[171] **Valk, J.** and **T. Alumae**, Voxlingua107: A dataset for spoken language recognition. 2021.

[172] **Vaswani, A.**, **N. Shazeer**, **N. Parmar**, **J. Uszkoreit**, **L. Jones**, **A. N. Gomez**, **L. Kaiser**, and **I. Polosukhin**, Attention is all you need. 2017. URL https://arxiv.org/pdf/1706.03762.pdf.

[173] **Vijayaditya Peddinti, S. K., Daniel Povey**, A time delay neural network architecture for efficient modeling of long temporal contexts. *In Interspeech*. 2015.

[174] **Villalba, J.**, **N. Chen**, **D. Snyder**, **D. Garcia-Romero**, **A. McCree**, **G. Sell**, **J. Borgstrom**, **L. P. García-Perera**, **F. Richardson**, **R. Dehak**, **P. A. Torres-Carrasquillo**, and **N. 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.

[175] **Villalba, J.** *et al.* (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.

[176] **Voxforge.org** (). Free and open source speech recognition engines (on linux, windows and mac). http://www.voxforge.org/. Accessed 06/25/2014.

[177] **Wang, X.**, **L. Li**, and **D. Wang**, Vae-based domain adaptation for speaker verification. *In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2019.

[178] **Wang, X.**, **J. Yamagishi**, **M. Todisco**, **H. Delgado**, **A. Nautsch**, **N. Evans**, **M. Sahidullah**, **V. Vestman**, **T. Kinnunen**, **K. A. Lee**, **L. Juvela**, **P. Alku**, **Y.-H. Peng**, **H.-T. Hwang**, **Y. Tsao**, **H.-M. Wang**, **S. L. Maguer**, **M. Becker**, **F. Henderson**, **R. Clark**, **Y. Zhang**, **Q. Wang**, **Y. Jia**, **K. Onuma**, **K. Mushika**, **T. Kaneda**, **Y. Jiang**, **L.-J. Liu**, **Y.-C. Wu**, **W.-C. Huang**, **T. Toda**, **K. Tanaka**, **H. Kameoka**, **I. Steiner**, **D. Matrouf**, **J.-F. Bonastre**, **A. Govender**, **S. Ronanki**, **J.-X. Zhang**, and **Z.-H. Ling** (2020). Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, **64**, 101114. ISSN 0885-2308. URL `https://www.sciencedirect.com/science/article/pii/S0885230820300474`.

[179] **Wu, Y.**, **C. Guo**, **H. Gao**, **X. Hou**, and **J. Xu**, Vector-Based Attentive Pooling for Text-Independent Speaker Verification. *In Proc. Interspeech 2020*. 2020. URL `http://dx.doi.org/10.21437/Interspeech.2020-1422`.

[180] **Wu, Y.-C.** and **W.-H. Liao**, Toward text-independent cross-lingual speaker recognition using english-mandarin-taiwanese dataset. *In 2020 25th International Conference on Pattern Recognition (ICPR)*. 2021.

[181] **Xia, W.**, **J. Huang**, and **J. H. Hansen**, 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)*. 2019.

[182] **Xia, W.**, **J. Huang**, and **J. H. Hansen**, 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)*. 2019.

[183] **Xia, W.**, **J. Huang**, and **J. H. L. Hansen**, Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019.

[184] **Xie, Q.** *et al.*, The multi-speaker multi-style voice cloning challenge 2021. *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[185] **Xie, Y.**, **H. Cheng**, **Y. Wang**, and **L. Ye** (2024). Domain generalization via aggregation and separation for audio deepfake detection. *IEEE Transactions on Information Forensics and Security*, **19**, 344–358.

[186] **Xu, J.**, **X. Tan**, **Y. Ren**, **T. Qin**, **J. Li**, **S. Zhao**, and **T.-Y. Liu**, Lrspeech: Extremely low-resource speech synthesis and recognition. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450379984. URL https://doi.org/10.1145/3394486.3403331.

[187] **Xu, J.**, **W. Zhou**, **Z. Fu**, **H. Zhou**, and **L. Li** (2021). A survey on green deep learning. *ArXiv*, **abs/2111.05193**. URL https://api.semanticscholar.org/CorpusID:243861089.

[188] **Xu, Y.**, **P. Terhörst**, **M. Pedersen**, and **K. Raja** (2024). Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, **5**(1), 93–106.

[189] **Yang, D.**, **D. Hovy**, **D. Jurgens**, and **B. Plank** (2025). Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, **51**, 689–703. URL https://aclanthology.org/2025.cl-2.10/.

[190] **Yang, S.**, **D. Das**, **J. Cho**, **H. Park**, and **S. Yun**, Domain Agnostic Few-shot Learning for Speaker Verification. *In Proc. Interspeech 2022*. 2022.

[191] **Yang, S.-W.**, **P.-H. Chi**, **Y.-S. Chuang**, **C.-I. J. Lai**, **K. Lakhotia**, **Y. Y. Lin**, **A. T. Liu**, **J. Shi**, **X. Chang**, **G.-T. Lin**, **T.-H. Huang**, **W.-C. Tseng**, **K. tik Lee**, **D.-R. Liu**, **Z. Huang**, **S. Dong**, **S.-W. Li**, **S. Watanabe**, **A. Mohamed**, and **H. yi Lee**, Superb: Speech processing universal performance benchmark. *In Interspeech 2021*. 2021. ISSN 2958-1796.

[192] **Yi, J.**, **R. Fu**, **J. Tao**, **S. Nie**, **H. Ma**, **C. Wang**, **T. Wang**, **Z. Tian**, **Y. Bai**, **C. Fan**, **S. Liang**, **S. Wang**, **S. Zhang**, **X. Yan**, **L. Xu**, **Z. Wen**, and **H. Li**, Add 2022: the first audio deep synthesis detection challenge. *In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022.

[193] **Yi, J.**, **J. Tao**, **R. Fu**, **X. Yan**, **C. Wang**, **T. Wang**, **C. Y. Zhang**, **X. Zhang**, **Y. Zhao**, **Y. Ren**, **L. Xu**, **J. Zhou**, **H. Gu**, **Z. Wen**, **S. Liang**, **Z. Lian**, **S. Nie**, and **H. Li**, Add 2023: the second audio deepfake detection challenge. *In DADA@IJCAI*. 2023. URL https://api.semanticscholar.org/CorpusID:258841572.

[194] **Yi, J.**, **C. Y. Zhang**, **J. Tao**, **C. Wang**, **X. Yan**, **Y. Ren**, **H. Gu**, and **J. Zhou** (2024). Add 2023: Towards audio deepfake detection and analysis in the wild. URL https://arxiv.org/abs/2408.04967.

[195] **Yousif, M.**, **J. J. Mathew**, **H. Pallan**, **A. S. Padda**, **S. D. Shah**, **S. Adamski**, **M. Reddiboina**, and **A. Pankajakshan** (2024). Enhancing generalization in audio deepfake detection: A neural collapse based sampling and training approach. URL https://arxiv.org/abs/2404.13008.

[196] **Yousif, M.**, **J. J. Mathew**, **H. Pallan**, **A. S. Padda**, **S. D. Shah**, **S. Adamski**, **M. Reddiboina**, and **A. Pankajakshan** (2024). Enhancing generalization in audio deepfake detection: A neural collapse based sampling and training approach. URL https://arxiv.org/abs/2404.13008.

[197] **Yu, Y.-Q.**, **L. Fan**, and **W.-J. Li**, Ensemble additive margin softmax for speaker verification. *In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[198] **Zeinali, H.**, **K.-A. Lee**, **J. Alam**, and **L. Burget** (2019). Short-duration speaker verification (sdsv) challenge 2020: the challenge evaluation plan. *ArXiv*, **abs/1912.06311**. URL `https://api.semanticscholar.org/CorpusID:209370807`.

[199] **Zen, H.** *et al.* (2019). Libritts: A corpus derived from Librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

[200] **Zhao, F.**, **H. Li**, and **X. Zhang**, A robust text-independent speaker verification method based on speech separation and deep speaker. *In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[201] **Zhao, Y.**, **W.-C. Huang**, **X. Tian**, **J. Yamagishi**, **R. K. Das**, **T. Kinnunen**, **Z. Ling**, and **T. Toda** (2020). Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*.

[202] **Zhou, Y.**, **X. Tian**, and **H. Li** (2021). Language agnostic speaker embedding for cross-lingual personalized speech generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3427–3439.

[203] **Zhu, D.** and **N. Chen** (2022). Multi-source domain adaptation and fusion for speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, 2103–2116.

[204] **Zhu, Y.**, **T. Ko**, **D. Snyder**, **B. Mak**, and **D. Povey**, Self-attentive speaker embeddings for text-independent speaker verification. *In Proc. Interspeech 2018*. 2018.

[205] **Zhu, Y.**, **S. Koppisetti**, **T. Tran**, and **G. Bharaj** (2024). Slim: Style-linguistics mismatch model for generalized audio deepfake detection. *ArXiv*, **abs/2407.18517**. URL `https://api.semanticscholar.org/CorpusID:271516320`.