BHASHINI

Indic Voice Technologies For An Inclusive Digital India

Toolkit for Developers

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

FAIR Forward

TRILEGAL

digital futures lab

ARTPARK
AI & Robotics Technology Park @ IISc

nasscom ai
(Industry Advisor)

Contents

Acknowledgements

This project would not have been possible without the contributions of many individuals and institutions. Our sincere gratitude to Bhashini, especially Mr Amitabh Nag, for his support. Many thanks to the GIZ team of FAIR Forward - Artificial Intelligence for All, a project funded by the German Ministry for Economic Cooperation and Development (BMZ).

We thank the Advisory Board and Working Group members for their time and valuable insights, which have informed this report. We thank Nasscom for its support as an industry advisor.

Our special thanks to the interview and workshop sprint participants whose perspectives helped refine our understanding of the voice-technology ecosystem in India.

The views and opinions expressed herein are exclusively those of the contributing authors, and do not represent the official views of any other organisation or entity associated with this project.

CONTRIBUTING AUTHORS

Artpark - Nihar Desai, Sujith P
Digital Futures Lab - Harleen Kaur, Dona Mathew
Trilegal - Jyotsna Jayaram, Padmavathi Prasad, Thomas Vallianeth

REVIEWERS
Aarushi Gupta, Aishwarya Salvi, Bhavika Nanawati, Philipp Olbrich

GUIDANCE
Jigar Doshi, Urvashi Aneja

COPYEDITING
Shreya Ramnath

LAYOUT AND DESIGN
Avani Tanya

Advisory Board and Working Group

ADVISORY BOARD

Amitabh Nag, CEO, Digital India Bhashini Division

Kalika Bali, Senior Principal Researcher, Microsoft Research India

Vinod Rajasekaran, Lead, Fractional CxO Project, Tech4Dev

Mitesh Khapra, Associate Professor, IIT Madras

Prasanta Kumar Ghosh, Associate Professor, IISc Bangalore

WORKING GROUP

Aaditeshwar Seth, Professor, IIT Delhi / Co-Founder, GramVaani

Brian DeRenzi, Head of Research and AI, Dimagi

Howard Lakougna, Senior Programme Officer, Gates Foundation

Jagadish Babu, COO, EkStep

Janki Nawale, Linguist, AI4Bharat

Pradeep Parappil, Co-Founder, Megdap

Soma Dhavala, IIT Jammu

Tahir Javed, AI4Bharat

Ujjwal Relan, Vice President, Samagra

Varun Hemachandran, Lead, OpenNyAI, Agami

Venkatesh Hariharan, India Representative, Open Invention Network

Vibhav Mithal, Associate Partner, Anand & Anand

Vineet Singh, CTO, Digital Green

Vivek Seshadri, Co-Founder, Karya

Industry Advisor

We acknowledge Nasscom for its role as an industry advisor, including its strategic inputs for outreach.

We particularly thank the following individuals:

Ankit Bose, Head of AI, Nasscom

M. Chockalingam, Technology Director, Nasscom

Shefali Mehra, Senior Associate, Nasscom

Kritika Oberoi, Associate, Nasscom

List of Abbreviations

ASR	Automatic Speech Recognition
AST	Automatic Speech Translation
BLEU	Bilingual Evaluation Understudy
BPCC	Bharat Parallel Corpus Collection
CER	Character Error Rate
DAPT	Domain-Adaptive Pretraining
DOI	Digital Object Identifier
DPDP Act, 2023	Digital Personal Data Protection Act, 2023
DPDP Rules	Digital Personal Data Protection Rules, 2025
DPIIT	Department for Promotion of Industry and Internal Trade
IT Act	Information Technology Act, 2000
IVR	Interactive Voice Response
LDCIL	Linguistic Data Consortium for Indian Languages
MLS	Multilingual LibriSpeech
NBFC	Non-Banking Financial Companies
NLP	Natural Language Processing
NLTM	National Language Translation Mission
OS	Open Source
PHCs	Primary Healthcare Centres
PI	Personal Information
Privacy Rules	Information Technology (Reasonable Security Practices and Sensitive Personal Data or Information) Rules, 2011
RAI	Responsible AI
SNR	Signal-to-Noise Ratio
SPI	Sensitive Personal Information
TDM	Text and Data Mining
TTS	Text-to-Speech
ULCA	Unified Language Contribution API
VAD	Voice Activity Detection
WER	Word Error Rate

Foreword

India’s digital transformation has been a remarkable phenomenon, impacting various sectors from finance to agriculture. However, in a country with such a linguistically diverse population and varying levels of digital literacy, digital technologies primarily centered around English still face significant accessibility challenges.

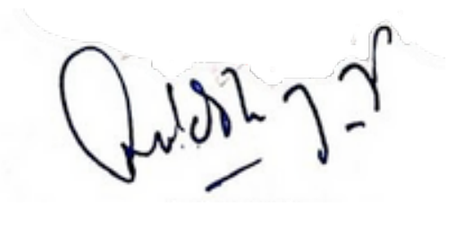
To address this gap in user interface design and last-mile connectivity, the Digital India Bhashini Division has been working to convene an intuitive, voice-based digital ecosystem in the country that serves the needs of millions who are currently excluded from digital services.

For a future with reliable, open-source and sustainable Indic Language AI, it is necessary for developers and policy makers to address the numerous challenges inherent to this space, including standardising language audio data collection, capturing linguistic nuances, and complying with legal and ethical guardrails.

By drawing on insights from linguistic experts, voice-technology developers, social impact organisations utilising voice technologies, and initiatives such as IndicVoices, SYSPIN, RESPIN and Project Vaani, the developers’ toolkit provides valuable perspectives on the practical challenges and potential technical solutions involved in building open-source voice technologies.

I encourage you to review the toolkit for a comprehensive understanding of the challenges and best practices within India’s voice-technology ecosystem. The insights offered are relevant not only for stakeholders in India but also for other contexts with limited resources and linguistic diversity. As language-centered digital technologies become integral to future digital infrastructure, initiatives like these are essential to understanding and advancing this important field.

We would be happy to receive your suggestions or feedback, if any, after you have reviewed the report.



Amitabh Nag,
Chief Executive Officer,
Digital India BHASHINI Division

Executive Summary

Executive Summary

The development of speech and language technologies in the Indian context is constrained not by a lack of innovation, but by persistent structural gaps in data representation, quality assurance, evaluation practices, and governance. Models trained on narrow or homogenised datasets risk underperforming for large segments of the population, while post-hoc ethical safeguards and deployment fixes are insufficient to address foundational exclusions embedded early in the development lifecycle.

This toolkit sets out a layered, lifecycle-oriented approach to building inclusive and robust speech artificial intelligence (AI) systems. It brings together strategies for diverse and representative data collection, linguistically informed model training, rigorous quality control, and deployment optimisation under real-world constraints, alongside embedded Responsible AI (RAI) practices.

1. Ensuring Diverse Representation:

a. **Develop diversity wishlist:** Leverage existing datasets and create a diversity wishlist based on demography, geography and linguistic nuances.

b. **Heterogeneous data collection:** Adopt a variety of data collection methods like crowdsourcing, field-based initiatives, and community media platforms, documenting different forms of speech through read, extempore and role play scenarios.

c. **Apply linguistic expertise:** To handle nuances like hybridism, code-switching, coarticulation variability, and morphological complexities, invite Indic language experts for inputs at the data collection stage.

d. **Use synthetic data** : Where feasible, use synthetic data to supplement gaps in data collection.

e. **Model training for linguistic nuances:** Account for linguistic nuances through pre-training and fine-tuning on code-switched databases and regular evaluation.

f. **Layered data strategy:** Use generic or foundational datasets and fine-tune with use-case specific datasets.

2. Enhancing Data Quality and Building Inclusive Applications:

a. **Implement quality control mechanisms:** Use rigorous quality control processes, including metadata verification (e.g., age, gender via video or WhatsApp calls), content checks (e.g., rejecting low-quality recordings based on error categories), and transcription accuracy assessments.

b. **Use detailed transcription guidelines:** Adopt two-level transcription guidelines: level 1 for verbatim transcription and level 2 for standardised transcription with tags for errors and linguistic features.

c. **Specialised tools for transcription:** Use specialised tools like Karya for data collection and Shoonya for transcription to ensure efficiency and scalability.

d. **Maintain Datacards**: They represent a paradigm shift toward responsible AI development, serving as comprehensive metadata documents that detail dataset creation, composition, and limitations.

e. **Better benchmarks for evaluation:** The disparities in datasets risk producing exclusionary outcomes, mainly when models are evaluated using a limited set of metrics like Word Error Rate (WER). Complementary metrics, such as answer error rate or intent accuracy, are necessary to reflect real-world usage better.

f. **Managing accent and pronunciation variations:** To generalise across accent and pronunciation variations, incorporate diverse speech datasets capturing regional differences. Regular fine-tuning on region-specific data and adversarial training to reduce accent bias improves generalisation. Voice cloning and accent correction techniques can be incorporated to tackle this challenge while building the solution.

g. **Managing noisy speech in datasets:** Noisy speech data, common in Indian settings, can be addressed by evaluating signal quality like Signal-to-Noise Ratio (SNR) and rejecting low-quality speech data. Employ noise-robust modelling techniques, such as synthetic noise augmentation or denoising autoencoders, as used in IndicWav2Vec and Indic Conformer.

h. **Model cards:** Complementing datacards, model cards serve as structured documents that provide essential context and

transparency for trained AI models.

i. **Model Deployment and Optimisation Strategies:** Optimised offline models and hybrid offline–online approaches enable reliable operation under limited connectivity by balancing local inference with cloud support when available. Speech-to-speech models further reduce end-to-end latency by removing intermediate processing steps."

3. **Embedding Responsible AI practices:**

a. **Meaningful engagement with language communities:** Data compilation process for speech and language technology must begin with meaningful engagement with associated language communities. This involves understanding their needs and aspirations regarding language technology and ensuring they have a say in how their data is used.

b. **Documentation and packaging:** Comprehensive documentation and standardised packaging including detailed guides on accessing, using, and contributing to the data and clear instructions for model deployment and fine-tuning.

c. **Ethical and Legal Pre-design:** The consent mechanism, personally identifiable information (PII) reduction protocols, parameters for RAI as applicable to the specific use case, and data storage boundaries must be designed upfront, not as an afterthought, to ensure the flywheel operates at scale and remains compliant with privacy laws and ownership boundaries.

d. **Obtain informed consent:** Where applicable, ensure that clear and unambiguous consent with an affirmative action is obtained from all participants, clearly explaining how their data will be used, stored, and shared.

e. **Protect privacy:** Implement robust privacy protections, ensuring compliance with data protection laws like the Digital Personal Data Protection Act, 2023 (DPDP Act) or General Data Protection Regulation (GDPR) for international collaborations. Anonymise all personal data where possible to limit the applicability of the DPDP Act or other data protection regimes globally. Privacy Enhancing Technologies (PETs) such as analysing voice patterns on the fly without storing PII, should be actively explored and supported by the ecosystem.

f. **Compliance with Copyright laws:** Under Indian law, particularly the Copyright Act, 1957, multiple layers of intellectual property protection may apply: to underlying text or transcripts (as literary works), to voice recordings (as sound recordings), and to curated metadata, provided each meets originality requirements. The use of any copyrighted information requires a license from the person who owns the copyright.

g. **Ensure transparency:** Be transparent about data collection purposes, share data management practices with participants and stakeholders.

h. **Foster accountability:** Establish mechanisms for accountability, such as regular audits and clear governance structures.

i. **Identifying the license under which open sourcing is carried out:** Licensing datasets and models is a critical decision that dictates their permissible use, modification, and distribution.

j. **Terms of use:** Adopt Terms of Use and Acceptable Use Policies that define how the data may be accessed, shared, modified, and redistributed. These policies should explicitly address privacy, consent, attribution, commercial use, and downstream redistribution to ensure legal compliance and ethical usage.

k. **Selecting a hosting platform:** The selection of a hosting platform for open source datasets and models should be guided by accessibility, scalability, version control, and community engagement features.

l. **Audit Techniques and Benchmarks:** Audit techniques and robust benchmarks evaluate system performance against specified safety and fairness criteria throughout the deployment lifecycle. For Indic languages, this would involve ongoing evaluation against diverse accent and dialect benchmarks to ensure equitable performance.

m. **Mitigating Misuse:** Regular and rigorous assessment of the AI system's safety properties, including its robustness to adversarial attacks and its resilience in unforeseen circumstances. This is particularly important for voice systems susceptible to audio manipulations or voice impersonation attacks. Appropriate mitigation mechanisms, such as not collecting PII, using appropriate licensing, and watermarks should be considered for open datasets and models.

1.
Building Voice Technology for India: Issues and Best Practices

1.Building Voice Technology for India: Issues and Best Practices

While smartphone access is expanding across the Global Majority, many users continue to face significant barriers to meaningful use.[1] Online content and applications remain text-heavy and complex, excluding over a billion people with limited literacy, most of whom live in low- and middle-income countries.[2] As access grows, the digital divide is therefore shifting from connectivity to quality of use, with marginalised groups disproportionately confined to narrow "application islands."[3] In this context, **voice technology presents a critical opportunity to overcome literacy barriers and enable inclusive digital participation**. Voice-based technologies are systems that process and respond to human voice commands. These technologies have transformed human-computer interactions, offering more intuitive modes of communication.

Gaps and Challenges in the Existing Ecosystem

AI-enabled voice technologies such as speech recognition, generation, translation, and transcription services are proliferating rapidly in India.[4] Despite promising developments, the ecosystem faces several key challenges that hinder the effective development, exchange, and use of open speech technologies in the country. These challenges contribute to a shortage of high-quality open speech datasets and models that are representative and inclusive, thereby slowing innovation in the open voice-technology space. This report aims to help developers identify and address the key challenges associated with open voice technologies in India.

1 B. Amland, A. Chine, S. Pandey, and A. Badshah, "Voice Recognition and Development Impact," Gates Open Research 7 (2023): 77, https://doi.org/.

2 Amland et al., 2023.

3 Application islands are easy to use applications that are widely used by marginalised groups, including women, due to the inability to adapt and apply skills to new applications. Source: GSMA, "Connected Women: The Mobile Gender Gap Report 2018" (GSM Association, 2018), https://www.gsma.com/.

4 For example, Shivam Saxena, "Why NBFCs are Betting Big on Vernacular Voice AI," Gnani.ai (2025), https://www.gnani.ai/.

About this Toolkit

This document focuses on the key challenges developers encounter and the practical approaches commonly adopted within India's voice-technology ecosystem to address them.

While we recognise that the field of voice technologies in India is rapidly evolving, we emphasise principles-based approaches that developers can adopt when designing their products, even as specific interventions continue to be replaced by newer ones. Broader structural issues and higher-level recommendations beyond the scope of developer action are addressed separately in **Building an Open and Responsible Voice Technology Ecosystem: Policy Recommendations for Digital Inclusion in India.**

In this toolkit, the term *"developers"* is used broadly to refer both to those building and curating open datasets and models for public use, as well as to downstream teams that rely on these resources to develop applications. While the focus is on contributors to the open voice-technology ecosystem, many of the practices outlined here are equally applicable to developers working outside the ecosystem, given the shared underlying principles and challenges.

India's voice-technology ecosystem faces a distinctive set of technical hurdles: extreme linguistic variation,[5] scarcity of high-quality data for many languages,[6] inconsistent annotation practices, and the reality of users interacting through low-cost devices in noisy, unpredictable environments. These challenges compound one another; models trained on narrow datasets struggle with accent drift, code-mixing, and spontaneous speech; poorly standardised pipelines introduce avoidable errors; and the absence of shared benchmarks and documentation hinders reuse and leads to repeated reinvention across projects.

This toolkit outlines the key technical problems developers are likely to encounter across the data, model, and deployment lifecycles, and presents concrete practices to address them. It examines how to design for linguistic and demographic diversity, ensure consistent and reliable annotation, manage noise and device variability, build robust evaluation protocols, and adopt documentation and governance standards that make datasets trustworthy and reusable. The aim is to provide builders with a realistic understanding of these challenges, along with practical guidance for building voice systems that remain accurate, inclusive, and dependable in India's real-world conditions.

5 Government of India, "Census 2011, Language" (2018), https://censusindia.gov.in
6 GIZ, "A Study on Open Voice Data in Indian Languages" (2020), https://www.bmz-digital.global

2. Ensuring Diverse Representation

2. Ensuring Diverse Representation

The gap in diversity and representation begins at the data collection stage and persists throughout the lifecycle of open voice technology development and dissemination. Underrepresentation of key demographic and contextual variables—such as geography, gender, age, socio-economic status, and ethnicity—can result in datasets that lack diversity.[7] The use of such datasets in model development can, in turn, contribute to reduced inclusivity in voice technologies.[8] Low-resource languages are particularly affected, as they often draw from limited and less diverse sources of data. Data may be mistranslated, contain nonsensical text scraped from the internet, or be limited to narrow domains such as religious texts and Wikipedia.[9] Moreover, available data often fails to adequately reflect how people actually speak and communicate, especially if the dataset is generated from scripted interactions or conversations between strangers, thereby ignoring colloquial usages.[10]

In this section, we highlight some key technical challenges and provide examples of best practices to mitigate them.

7 GIZ, 2020.

8 Yvonne Kamegne, Eric Owusu, and Helina Oladapo, “Alexa, Why Can’t You Hear Our Accents: Cross Cultural Studies on the Inclusivity of Voice Recognition Systems,” In Social Computing and Social Media: 17th International Conference, SCSM 2025, Held as Part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings, Part II, 62–71. Berlin and Heidelberg: Springer-Verlag, 2025. https://doi.org/.

9 Gabriel Nicholas and Aliya Bhatia, “Lost in Translation: Large Language Models in Non-English Content Analysis,” arXiv pre print arXiv:2306.07377 (2023), https://arxiv.org

10 Interview with X, virtual, 20 June 2025.

2.1 Challenges in Ensuring Diverse Representation

a. **Data scarcity:** There is limited publicly available data for low-resource languages.[11] Since machine learning models depend on both quantity and quality of data, insufficient data often results in lower performance.[12] Even secondary data sources are limited. For example, in the case of text data, extensive web crawling still yields far less than what modern Large Language Models (LLMs) require. Most high-quality content remains locked in physical libraries, making national-scale, machine-readable digitisation essential rather than simple scanning. Practical constraints in implementing effective data flywheel mechanisms within open-source initiatives result in less representative datasets compared to closed models.[13] Data scarcity negatively impacts the quality and coverage of benchmarks, leading to a mismatch between reported development performance and real-world performance.

b. **Lack of standardised script:** Primarily oral languages that lack standardised written scripts[14] are often invisibilised as they do not fit into conventional orthographic taxonomies. For example, in their work on the Gormati language, Reitmaer et al. (2024) found that verbatim translation approaches used for written languages are ineffective for oral languages.[15] While written languages depend heavily on documenting conversations and stories, speakers of oral languages rely on memory-based practices for recalling information and do not need visual aids.[16] In Project Vaani,[17] researchers working on languages such as Chakma and Garo from the North-Eastern states in India faced challenges locating documents to support the script of these languages. Some Indian languages, such as Manipuri use multiple scripts. Different communities often prefer or advocate for different scripts, which makes

11 Wushour Slam, Yanan Li, and Nurmamet Urouvas, “Frontier Research on Low-Resource Speech Recognition Technology.” Sensors 23, no. 22 (2023): 9096. https://www.mdpi.com/.
12 Mohammed et. al., “The Effects of Data Quality on Machine Learning Performance on Tabular Data,” arXiv preprint arXiv:2207.14529, https://arxiv.org/.
13 Interview with Janki Nawale, Linguist / Researcher at AI4Bharat on 25/07/2025.
14 Neeru Misra, “Indian Languages Without Script Must be Saved from Extinction,” The Sunday Guardian Live, December 29, 2024, https://latest.sundayguardianlive.com/.
15 Thomas Reitmaier, Dani Kalarikalayil Raju, Ondrej Klejch, et al., “Cultivating Spoken Language Technologies for Unwritten Languages,” in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (New York: ACM, 2024), 1-17, https://dl.acm.org/.
16 Reitmaier et al., 2024.
17 VAANI Team, “VAANI: Capturing the Language Landscape for an Inclusive Digital India,” VAANI, https://vaani.iisc.ac.in/. Project Vaani is a collaborative initiative by the Indian Institute of Science (IISc), ARTPARK, and Google to develop an inclusive Indic multimodal dataset.

standardisation difficult and, in some cases, politically sensitive.[18]

c. **Logistical challenges:** Low-resource languages are often native to rural areas and remote regions, where infrastructural constraints pose significant challenges.[19] Many rural and tribal communities lack reliable internet connectivity and consistent electricity, hindering the use of digital recording tools and the efficient transfer of large audio files. Poor road networks and limited transportation options further exacerbate these difficulties, making physical access to communities time-consuming, expensive, and logistically complex.

d. **Challenges in recruiting participants:** Recruitment and engagement become arduous due to lower awareness of field-based initiatives, initial mistrust of external teams, and the need for culturally sensitive approaches to explain consent and build rapport, especially given varying literacy levels across different data contributors.[20] Mobilising support to reach speakers of diverse languages and dialects, identifying volunteers, and coordinating these efforts is time-consuming and resource-intensive, often requiring collaboration with grassroots organisations, linguistic departments within universities, and others.

e. **Selection bias in data sampling:** Data collection choices further drive gaps in available data. For example, the Switchboard corpus predominantly includes young, educated native English speakers as participants were largely recruited through universities and researchers' contacts.[21] Attitudes and ideologies about what counts as "good" English may further discourage participation from speakers of marginalised communities.[22] This phenomenon is also noted in Indian language datasets, where speech data providers are primarily from urban areas with graduate degrees.[23] Crowdsourcing platforms also assume access to a computer and the internet, which further skews participation. Reliance on predominantly urban contributors leads to a lack of diversity in the collected dataset. It is important to note that even within the same language, dialects spoken by rural Indians can differ from those spoken by urban Indians. Consequently, such non-diverse datasets are insufficient for building technologies that effectively serve rural users.[24]

18 Interview with XX on 25/07/2025.
19 Pratik Joshi et al., "Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities," arXiv preprint arXiv:1912.03457 (2019), https://aclanthology.org/.
20 Similar challenges are seen in participatory research including healthcare research. For example, see Ava Reck et al., "Building Trust in Rural Communities: Recruitment and Retention Strategies in Developmental Science," Frontiers in Public Health 13 (2025): 1586988, https://pmc.ncbi.nlm.nih.gov/.
21 Nina Markl, "Mind the Data Gap(s): Investigating Power in Speech and Language Datasets," in 2nd Workshop on Language Technology for Equality, Diversity, Inclusion 2022 (Association for Computational Linguistics, 2022), 1-12, https://aclanthology.org/.
22 Markl, "Mind the Data Gap(s)," 2022.
23 Basil Abraham et al., "Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers," in Proceedings of the Twelfth Language Resources and Evaluation Conference (2020), 2819-2826, https://aclanthology.org/.
24 Abraham et al., 2020.

f. **Linguistic complexities:** India is home to an exceptionally diverse linguistic landscape, with over 19,500 mother tongues, including languages and dialects, [25] each carrying unique linguistic features and cultural context. This diversity introduces significant nuance and complexity into data collection processes. The English-centric design of preprocessing techniques (e.g., tokenisation) and language models often fails to account for the linguistic diversity, morphological complexity, and dynamic evolution of languages through code-mixing[26] and code-switching[27], practices often absent in English.[28] Moreover, the extensive inflectional and agglutinative features of Indian languages result in greater vocabulary sizes, presenting challenges for model integration. In Table 1 below, we examine some of the linguistic properties of Indic languages relevant for data collectors and curators.

Although the challenges of representativeness and linguistic complexity are usually closely intertwined, they represent distinct concerns that require differentiated approaches.

TABLE 1: Common Linguistic Nuances in Indic Languages

FEATURE	MEANING
Code-mixing and code-switching	Both are features of mixed language. While code-mixing involves mixing words,[29] mixing sentences is code-switching.[30]
Coarticulation variability	Variation in articulation occurs due to surrounding speech segments.[31] In tonal languages like Mizo and Manipuri, pitch and tone can change the meaning of words, so speech data must be carefully collected and labelled to capture these differences correctly.[32]

25 Government of India, "Census 2011, Language" (2018), https://censusindia.gov.in/
26 Alternating between words and phrases from multiple languages in the same sentence.
27 Alternating words and phrases from multiple languages between sentences and clauses.
28 Farhana Shahid, Mona Elswah, and Aditya Vashistha, "Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages," arXiv preprint arXiv:2501.13836 (2025), https://arxiv.org/.
29 Pratibha, "Expanding Research Horizons for Hinglish Text by Tackling Challenges and Research Gaps." Journal of Information Systems Engineering and Management 10 (2025): 481-497, https://www.researchgate.net/.
30 Suneeta Thomas, "Code-Switching in Spoken Indian English: A Case Study of Sociopolitical Talk," Journal of Contemporary Philology 4, no. 1 (2021): 7-41.
31 Daniel Recasens, "Lingual Coarticulation," in Coarticulation: Theory, Data and Techniques, ed. William J. Hardcastle and Nigel Hewlett (Oxford: Oxford University Press, 1999), 80-104, https://oxfordre.com/.
32 N. John Kuotsu, "Advancing Natural Language Processing for Underrepresented Tibeto-Burman Languages in Northeast India," Sch J Eng Tech 12 (2024): 342-348, https://www.saspublishers.com/.

Morphological patterns	Words change their form to convey different meanings, such as tense, number, gender, or case. Languages in the Tibeto-Burman family are characterised by complex morphological patterns.[33]
Agglutination	Agglutinative languages contain words that are formed by combining multiple morphemes, each carrying a distinct semantic or syntactic meaning.
Pitch, accent and dialectical variation	Differences in pitch, accent, and dialect can change the meaning of words.
Syntactic variation	Differences in sentence structure and word order can confuse speech systems if not well represented in low-resource language data. For example, South Asian languages like the Indo-Aryan and Dravidian language families follow subject-object-verb, with some variations.[34] Therefore, the rules of English linguistics for Automatic Speech Recognition (ASR) or translation cannot be applied effectively for these languages.

g. **Limited Representation of Indic Languages in Modern Open-Source AI Models:** Open-source models released on the global stage often have limited or no coverage of Indic languages.[35] [36] This lack of support significantly limits their applicability for Indic voice-based applications, where accurate understanding of local languages is essential.

h. **Performance Limitations for Narrow Use Cases:** Most available models are generic in nature and often perform poorly when applied to specific domains or specialised use cases, largely due to the absence of domain-specific data in the training process.

33 Kuotsu, 2024.
34 Veneeta Dayal and Anoop Mahajan (eds)., “Clause Structure in South Asian Languages: General Introduction,” in Clause Structure in South Asian Languages (Dordrecht: Kluwer Academic Publishers, 2005), 1–11, https://bpb-us-w2.wpmucdn.com/.
35 Monica Sekoyan et. al., “Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST,” arXiv preprint arXiv:2509.14128 (2025), https://arxiv.org/
36 Abdelrahman Abouelenin et al., “Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs,” arXiv preprint, March 2025, https://doi.org/.

2.2 Best Practices for Ensuring Diverse Representation

We present below key practices that developers can adopt to overcome the challenges highlighted above:

a. **Develop a diversity wishlist:** A diversity wishlist is an approach in which entities engaging in data collection list the attributes they want to include in their dataset to enhance diversity. These include features such as data principal[37] age, geographical coverage, and the type of language or dialect to be collected. By defining the diversity wishlist at the outset, developers can ensure that future users of the open dataset are aware of the strengths and limitations of the coverage of data with respect to their specific use case. This, in turn, enables the developers to prepare a custom plan for fine-tuning if necessary.

   While diversity wishlists can be customised to the requirements of the data-collection team, it is important to include perspectives from experts like linguists when determining what should be prioritised for diversity wishlists. This helps avoid blind spots that may arise from a technical framing of diversity. SYSPIN[38] and RESPIN[39,] one of the earliest efforts in the sector, targeted nine Indian languages across 40 districts. Other larger initiatives, such as the IndicVoices dataset, targeted 22 Indian languages across 408 districts, setting an example for broad demographic and geographic coverage.[40] The diversity plan for the IndicVoices database included representation across demographics, including conditions related to age (e.g.,15% representation from age brackets 18–30, 30–45, 45–60, 60+); equal representation across gender, educational levels, professions, locations (including 60–80% of districts where the language is primary), content domains (e.g., legal, governance, health), speech styles (read, extempore, conversational), vocabulary, downstream usage (e.g., digital payments, government services), and recording conditions (e.g., noisy environments, various devices).

37 A data principal is a person whose data is being collected, used, or processed.
38 Abhayjeet et al. 'SYSPIN_S1.0 Corpus - A TTS Corpus of 900+ hours in nine Indian Languages', 2025.
39 Saurabh Kumar et. al., "RESPIN-S1.0: A Read Speech Corpus of 10000+ Hours in Dialects of Nine Indian Languages," In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*, 2025, https://openreview.net/.
40 Tahir Javed et al., "IndicVoices: Towards Building an Inclusive Multilingual Speech Dataset for Indian Languages," arXiv preprint arXiv:2403.01926 (2024).

2. ENSURING DIVERSE REPRESENTATION

While most datasets are language-specific, an alternative approach to enhance inclusivity across languages is to adopt a geography-specific (geocentric) data collection approach. Researchers working on SYSPIN and RESPIN identified the possible drawbacks of language-specific collection and tested these learnings in Project Vaani.

Prioritising low-resourced languages

For low-resource languages where baseline corpora are absent, initial collection efforts must prioritise use-case-specific data to enable rapid bootstrapping and demonstrate immediate community benefit, which can, in turn, attract the necessary funding and resources.

The table below compares both these approaches:

LANGUAGE-BASED	GEOCENTRIC
Tailored to individual languages	Centralised, often more uniform across regions
Easy to execute on a smaller scale	May involve a complex collection process
Remote collection is possible with quality control	Remote collection may not yield desired results
Collection is possible from secondary sources	Difficult to collect from secondary sources
May miss linguistic and dialectal variations	Better guarantees on language and regional coverage
Can overlook language diversity	More likely to ensure language diversity
Suitable for building models for widely spoken languages in a short time	Essential for building inclusive voice technology with broad coverage

Each approach has pros and cons. The downstream applications' specific requirements should guide the choice of the data collection strategy.

Downstream developers should consider targeted data collection tailored to their specific use cases, particularly when existing open datasets are limited. Developers are encouraged to follow data collection specifications released as part of large-scale initiatives such as IndicVoices, RESPIN, SYSPIN, and VAANI. In addition, developers may consider the use of synthetic data generation techniques for domain adaptation.[41]

Leveraging existing datasets

Despite the shortage of representative datasets for Indic languages, downstream developers can leverage several foundational datasets instead of collecting data from scratch. These resources can be used in conjunction with targeted, use-case-specific data collection:

- **RESPIN**[42]**:** A read speech corpus of 10,000+ hours in dialects of nine Indian languages hours across nine languages
- **Common Voice**[43]**:** 35,921 hours across 286 languages, including more than 13 Indic languages[44]
- **IndicVoices**[45]**:** 19,550 hours across 22 languages
- **Vaani**[46]**:** 31,255 hours across 109 languages

Through preprocessing and fine-tuning, these datasets can be adapted to specific needs. For instance, Common Voice's diverse samples can address demographic underrepresentation, enhancing dataset inclusivity.

Additionally, unique datasets like SYSPIN (a 720-hour TTS dataset covering nine languages) offer access to high-quality TTS datasets in a resource-starved environment. Developers should verify signal properties, transcription structure, recording conditions, and intended use case, and perform the necessary preprocessing of these datasets to ensure alignment with their use case.

41 Minh Tran et. al., " A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic Data," arXiv pre print arXiv:2403.01926 (2024), https://arxiv.org.
42 "About RESPIN," RESPIN, https://respin.iisc.ac.in/about.
43 Common Voice 23.0 Live on Mozilla Data Collective," Mozilla Data Collective, https://community.mozilladatacollective.com/.
44 "Datasets," Mozilla Data Collective, https://datacollective.mozillafoundation.org/datasets.
45 Javed et al., "IndicVoices."
46 Artpark, "Capturing the Language Landscape for an Inclusive Digital India," 2025, https://vaani.iisc.ac.in/

b. **Use heterogeneous data collection methods:** Developers should combine crowdsourcing, field-based initiatives, and community media platforms to reach diverse populations, including rural and low-resource language speakers, thereby increasing the diversity of language sources. Platforms like BhashaDaan[47] and field efforts like Project Vaani and GramVaani[48] demonstrate how participatory and grassroots approaches can enhance inclusivity. An appropriate collection strategy should be chosen by carefully balancing factors such as cost, complexity, data authenticity, and metadata accuracy requirements. Achieving this balance is critical for designing a data collection pipeline that is both effective and scalable.

Collaboration for heterogeneous data collection

Collaborate with local organisations, universities, and community leaders to mobilise participants from diverse backgrounds. Provide clear project information and obtain informed consent to build trust and ensure ethical participation. The IndicVoices project partnered with data collection agencies, foundations, and universities across 408 districts to achieve the goals of their diversity plan.[49] In Project Vaani, researchers reported leveraging Primary Healthcare Centres (PHCs) as a nodal point to connect with participants.

In addition, developers can consider collecting data from varied sources, including read speech (e.g., from Wikipedia articles), extempore speech (e.g., prompted questions), and conversations (e.g., role-play scenarios), to capture natural language use. The Indic TTS project used multiple text sources like newspapers and blogs to ensure more diverse coverage.[50]

c. **Streamline linguistic nuances:** Organisations involved in data collection and curation, such as IndicVoices and Project Vaani, engage linguists and native speakers to handle nuances like code-mixing, code-switching, coarticulation variability, and morphological complexities. In addition, downstream developers use mechanisms that are mindful of linguistic nuance for better performing applications:

- Language experts are essential to identify strategies for data collection, including considerations around linguistic nuance,

47 Bhashini, “Bhasha Daan Empower India’s Linguistic Diversity,” 2025, https://bhashini.gov.in.
48 Mobile Vaani Connect Network, Gram Vaani, https://gramvaani.org/.
49 Tahir et al., “Indicvoices.”
50 CIS, “Making Voices Heard: Policy Brief,” 2022, https://voice.cis-india.org/.

accents, dialects, and sampling mechanisms. Engaging with **language experts ensures accurate representation of linguistic features, particularly for languages like Chakma and Garo, which lack standardised scripts.** For non-standardised scripts, developing orthographies in collaboration with linguists and integrating them into datasets like IndicOOV can be particularly useful.[51]

- Developing **standardised data collection protocols that account for variations in scripts, orthography, and phonetics helps scale up the data collection process**. Tools like the Karya app, modified to display localised prompts and hints in 22 Indian languages, facilitate this process.[52]

- From a model development perspective, **the lack of script standardisation necessitates the use of script-agnostic models trained on datasets** like IndicCorp, covering multiple scripts (e.g., Devanagari, Tamil, Bengali).[53]

- Approaches like **text normalisation and script unification may help mitigate the challenges of non-standardised scripts.**[54]

- Tools like **iNLTK**,[55] an Indic language counterpart inspired by libraries such as spaCy and NLTK, are being developed to **improve the limited NLP infrastructure** for Indian languages.

- Adopt **standardised benchmarks**, such as Vistaar and IndicSUPERB, to evaluate datasets and models while considering linguistic nuances.[56]

d. **Synthetic data:** Synthetic data cannot substitute genuine human recordings, as its quality is inherently limited by the capabilities of the base TTS system. However, it can serve as a valuable complementary resource to generate specific named entities, rare words, or domain concepts for targeted use-case adaptation. Developers should consider employing data augmentation techniques—such as synthetic voice generation, as used in Bhasanuvaad's 44,400-hour dataset—to bolster low-resource language data.[57] In addition, transfer learning from high-resource to low-resource languages can

51 Javed et al., "IndicVoices"
52 DAIA Tech Pvt Ltd, Karya Application, https://play.google.com/.
53 Anoop Kunchukuttan et al., "Ai4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages," arXiv preprint arXiv:2005.00085 (2020), https://arxiv.org/.
54 Bhavyajeet Singh, Pavan Kandru, Anubhav Sharma, and Vasudeva Varma, "Massively Multilingual Language Models for Cross-Lingual Fact Extraction from Low Resource Indian Languages," arXiv preprint arXiv:2302.04790 (2023).
55 Gaurav Arora, "inltk: Natural Language Toolkit for Indic Languages, arXiv preprint arXiv:2009.12534 (2020).
56 "Leveling Up NLP4 Indian Langs," RBCDSAI IIT Madras, https://rbcdsai.iitm.ac.in/.
57 BhasaAnuvaad, "BhasaAnuvaad: A Speech Translation Dataset for 14 Indian Languages," https://arxiv.org/.

substantially enhance performance, as demonstrated by IndicTrans2, which improves translation for languages like Santali by pretraining on Hindi.[58] These approaches ensure equitable model performance across India's linguistic spectrum.

e. **Model training for code-switching:** Code-switching, common in Indian multilingual contexts, requires models to effectively process mixed-language utterances. Multilingual pretraining with datasets like IndicCorpora, which include code-switched text and speech, can be implemented to train models like IndicBERT and IndicBART[59]. Fine-tuning on code-switched datasets, such as those from IndicVoices and Vaani, reduces transcription errors and enhances user experience in applications like voice assistants[60]. Finally, regular evaluation with benchmarks like MUCS[61], which include code-switched scenarios, ensures robustness.[62]

f. **Layered data strategy:** Model developers should adopt a layered data strategy. For general-purpose systems, data can be 70–80% generic—covering diverse demographics, accents, and dialects—and 20–30% domain-specific to enhance named entity and vocabulary recognition (e.g., medical or financial terms).[63]

Downstream developers should curate specialised corpora to address the underrepresentation of domain-specific vocabulary in sectors like agriculture, healthcare, and education. For instance, speech data capturing medical terminology in Hindi or agricultural terms in Tamil can be collected and integrated into datasets like IndicVoices. Researchers from ARTPARK & ARMMAN, working on an LLM-based assistant for public health workers, report using this approach. Some datasets, such as RESPIN, are already domain-focused (agriculture and finance).

In addition, developers may consider employing domain-adaptive pretraining (DAPT) to fine-tune models like IndicWav2Vec, thereby enhancing recognition of context-specific expressions.[64] This improves performance in applications like healthcare chatbots and educational tools. Adding synthetic data may further support these efforts.

58 Jay Gala et al., "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages," arXiv preprint arXiv:2305.16307 (2023), https://arxiv.org/abs/2305.16307

59 M. C. S. Priya, D. K. Renuka, L. A. Kumar, et al., "Multilingual Low Resource Indian Language Speech Recognition and Spell Correction Using Indic BERT," Sādhanā 47 (2022): 227, https://doi.org/10.1007/s12046-022-01973-5.

60 Javed et al., "IndicVoices."

61 Anuj Diwan et. al., "MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," in Proceedings of Interspeech 2021, 2446–50, https://www.isca-archive.org/interspeech_2021/diwan21_interspeech.html

62 Please see, AI4Bharat, "AI4Bharat," https://ai4bharat.iitm.ac.in/areas/asr

63 Inputs from workshop dated November 7, 2025.

64 Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra, "Towards Building ASR Systems for the Next Billion Users," In Pro ceedings of the AAAI conference on Artificial Intelligence, 36, no. 10 (2022), 10813-10821, https://arxiv.org/.

3. Enhancing Data Quality and Building Inclusive Applications

3. Enhancing Data Quality and Building Inclusive Applications

The quality and usability of open source speech datasets are critical to ensuring that the downstream voice-based tools function effectively across diverse real-world contexts. Poor-quality data can result in tools that are difficult to use, less inclusive, and potentially harmful. Below, we outline key challenges affecting the quality and usability of voice datasets and the tools built on them.

3.1 Challenges in Ensuring Data Quality and Building Inclusive Applications

a. **Speech naturalness and intelligibility:** Humans prefer natural-sounding voices in their interpersonal interactions, as they are perceived as more socially desirable. By contrast, unnatural voices may sound nasal or robotic, or may differ from regular human speech in pitch contour, temporal structure, or spectral composition.[65]

   There are many ways in which datasets collected may be distorted or unnatural. For instance, issues like multi-speaker overlap, background noise, volume inconsistencies, or poor articulation affect the intelligibility of the speech and reduce the dataset's value for training usable models.[66]

b. **Poor annotation and labelling:** Inaccurate or inconsistent labelling, errors in transcriptions, incorrect speaker tags, or misclassified language can introduce noise into datasets, adversely affecting model performance, compromising the effectiveness of downstream applications, and limiting reliable model comparison.[67] These challenges are exacerbated by the shortage of qualified transcribers who are both native language speakers and proficient writers familiar with the spellings and grammatical structures of these languages. Consequently, subjectivity in transcription is common, leading to inconsistent data.

c. **Lack of data collection and quality check tools:** The development of data curation tools has primarily focused on Western languages, particularly English. This Western-centric bias fails to address the linguistic diversity, cultural nuances, and evolving language practices of the Global Majority.[68]For instance, dependency parsers, which identify grammatical relationships between words, are unavailable

65 Christine Nussbaum, Sascha Frühholz, and Stefan R. Schweinberger, "Understanding voice naturalness." Trends in Cognitive Sciences (2025), https://www.sciencedirect.com/.
66 Ali Sartaz Khan, Tolulope Ogunremi, Ahmed Adel Attia, and Dorottya Demszky, "Multi-Stage Speaker Diarization for Noisy Classrooms," arXiv preprint arXiv:2505.10879 (2025), https://arxiv.org/.
67 Label Studio, "6 Costly Data Labeling Mistakes and How To Avoid Them," September 2022, https://labelstud.io/
68 Pratik Joshi, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in Proceedings of ACL 2020, 2020, https://aclanthology.org/

for many Tibeto-Burman languages,[69] and high-quality Voice Activity Detection (VAD) systems remain scarce for most Indic languages. The absence of basic tools, such as keyboards and spell-check systems tailored for Indic scripts, further hampers scaled data collection efforts for low-resource languages.[70]

d. **Lack of standardised dataset evaluation practices:** Existing dataset evaluation frameworks are heavily skewed toward Western languages. Researchers argue that metrics like Word Error Rate (WER) and Character Error Rate (CER) are inadequate for evaluating dataset quality for Indian languages, as they fail to account for linguistic complexities like code-switching, tonal variations, and cultural nuances.[71] The absence of appropriate benchmarks limits the ability to assess dataset quality across India's diverse linguistic landscape.

e. **Performance disparities and bias risks:** Voice models often exhibit uneven performance across languages, accents,[72] and speaker demographics due to data gaps and under-representation in training datasets.[73] For instance, users with regional or rural accents may experience lower recognition accuracy than urban English speakers. Utilising unbalanced datasets may result in models that are insufficiently culture-aware.[74]

For Automatic Speech Translation (AST) in particular, there remains a significant execution gap for Indian languages, primarily due to the lack of a specific translation dataset that includes audio for training and evaluating models.[75]

In addition, deployments in dialect-rich states prioritise monolingual dialectal accuracy over multilingual efficiency. Monolingual models typically offer better performance, lower size, and faster inference, while multilingual models require larger footprints and higher costs, despite offering unified service.[76]

69 N. John Kuotsu, "Advancing Natural Language Processing for Underrepresented Tibeto-Burman Languages in Northeast India," Sch J Eng Tech 12 (2024): 342-348, https://www.saspublishers.com/.

70 Interview with Aaditeshwar Seth, Professor, Department of Computer Science and Engineering, IIT Delh, virtual, July 21, 2025.

71 Anushka Singh, Ananya B. Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M. Khapra, "How Good is Zero-Shot MT Evaluation for Low Resource Indian Languages?," arXiv preprint arXiv:2406.03893 (2024), https://arxiv.org/.

72 Sarah Jassim and Abdulmohsin, Husam, Accent Classification Using Machine Learning Techniques: A Review," International Journal of Computer Information Systems and Industrial Management Applications, 17 (2025): 421-451. 10.70917/ijcisim-2025-0028 https://www.researchgate.net/.

73 Tanvina Patel, Wiebke Hutiri, Aaron Yi Ding, and Odette Scharenborg, "How to Evaluate Automatic Speech Recognition: Comparing Different Performance and Bias Measures," arXiv preprint arXiv:2507.05885 (2025), https://arxiv.org/.

74 Culture-awareness is the ability of LLMs and NLP systems to understand the context in which they are asked to perform a particular task and how that context varies with culture. In Pawar et al., "Survey of Cultural Awareness in Language Models: Text and Beyond," Computational Linguistics (uncorrected) (2025), https://doi.org/

75 Sanket Shah, Kavya Ranjan Saxena, Kancharana Manideep Bharadwaj, Sharath Adavanne, and Nagaraj Adiga, "IndicST: Indian Multilingual Translation Corpus For Evaluating Speech Large Language Models," in 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), (IEEE, 2025), 1-5, https://ai-labs.olakrutrim.com/.

76 Input from workshop dated November 7, 2025.

Limited coverage of Indic languages in state-of-the-art speech models is another significant limitation.[77] [78]

f. **Infrastructure challenges:** Limited mobile network coverage, low-end edge devices, high inference costs, and end-to-end latency remain significant challenges. In addition, India's telephony infrastructure must evolve to better handle low-quality networks while enabling modern, reliable, and low-latency communication channels. Compute infrastructure for model training and inference in India is also limited. At a national level, the available compute capacity is orders of magnitude smaller than that operated by a single organisation like Meta.[79]

77 Monica Sekoyan et al, "Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST," arXiv preprint, September 2025, https://doi.org/.
78 Zhiliang Peng, et al., "VIBEVOICE Technical Report," arXiv preprint, August 2025, https://doi.org/.
79 Interview with XX on 25/07/2025.

3.2 Best Practices for Quality and Building Inclusive Applications

a. **Implement quality control mechanisms:** Use rigorous quality control processes, including metadata verification (e.g., age, gender via video or WhatsApp calls), content checks (such as rejecting low-quality recordings based on error categories), and transcription accuracy assessments.[80] IndicVoices implemented such checks to ensure dataset reliability.

b. **Use detailed transcription guidelines:** Adopt a two-level transcription framework: level 1 for verbatim transcription and level 2 for standardised transcription with tags for errors and linguistic features (e.g., [baby_crying], [code-switching]). Tools like Shoonya support efficient transcription workflows with a maker-superchecker process.[81] In addition, maintain comprehensive documentation of linguistic variation and provide clear guidelines for transcribers to handle complex features like agglutination, syntactic variations, and tonal differences. This helps ensure that the data collected is error-free and standardised, while also enabling downstream users to more easily reuse the datasets and reduce the time and effort required for data processing prior to usage.

To address the shortage of trained labelling and annotation talent, invest in recruiting new trainees and support their skill over an extended period. Researchers in Project Vaani report having recruited and trained more than 3000 native speakers from approximately 160 districts for this purpose.

c. **Specialised tools for transcription:** Use specialised tools like Karya for data collection and Shoonya for transcription to ensure efficiency and scalability.[82] These tools should be user-friendly and accessible to participants with varying levels of digital literacy.

d. **Datacards:** Datacards represent a paradigm shift toward responsible AI development, serving as comprehensive metadata documents

80 Javed et al., “Indicvoices.”
81 AI4Bharat, “Shoonya,” https://ai4bharat.iitm.ac.in/tools/Shoonya
82 DAIA Tech Pvt Ltd, Karya Application, https://play.google.com/ and AI4Bharat, “Shoonya”, https://ai4bharat.iitm.ac.in/

that detail dataset creation, composition, and limitations. As noted by Pushkarna et al. (2022),[83] such documentation is essential for fostering transparency and accountability in machine learning systems.

Essential datacard elements include:

i. **Origins and collection methods:** Documentation of data gathering methodologies, including crowdsourcing platforms (e.g., Mozilla Common Voice), professional recordings, or synthetic generation. For Indic language datasets, this would include details on the sourcing of parallel corpora and native speaker contributions.

ii. **Speaker demographics:** A comprehensive breakdown, including age distribution, gender representation, native language, regional accents, and socioeconomic backgrounds. This is particularly vital for Indian languages, which exhibit significant accentual and dialectal diversity, as highlighted by projects like Svarah[84] for Indian English accents and LAHAJA[85] for Hindi accents.

iii. **Recording conditions:** Technical specifications such as recording environment, equipment used, noise levels (SNR ratios), and audio quality metrics help determine whether a dataset meets the quality requirements for a given application and informs appropriate preprocessing, filtering, and model selection strategies tailored to each use case.

iv. **Ethical considerations:** Documentation of informed consent protocols, data anonymisation procedures, and bias mitigation strategies, including clarity on how data from vulnerable populations or specific regional groups is handled.

v. **Language and dialect coverage:** Detailed mapping of included languages, dialects, and regional variations. For instance, datasets like Bharat Parallel Corpus Collection (BPCC) document their coverage of all 22 scheduled Indian languages.

83 Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson, “Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI,” in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, arXiv preprint, April 2022,1776–1826.

84 Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra, “Svarah: Evaluating English ASR Systems on Indian Accents,” in Proceedings of Interspeech, (Dublin, Ireland, August 2023): 5230-34, https://doi.org/.

85 AI4Bharat, Lahaja, GitHub, https://github.com/.

vi. **Phonetic and prosodic details:** Information on tone, pitch, intonation, stress patterns, and other acoustic features is critical to capture the nuances of Indic languages.

vii. **Legal considerations:** Information on the kinds of data collected (i.e., whether it constitutes any personally identifiable information (PII), anonymised data or pseudonymised data, etc.), dataset maintenance considerations, and end-use limitations, to name a few.

e. **Better benchmarks for evaluation:** Disparities in datasets risk producing exclusionary outcomes, particularly when models are evaluated using a narrow set of metrics like Word Error Rate (WER). Complementary metrics—such as answer error rate[86] or intent accuracy— are necessary to better reflect real-world usage.

   Developers should design evaluation metrics that account for code-switching, tonal variations, and cultural expressions, as highlighted in BhasaAnuvaad. Complement automated metrics with human-centric evaluations to ensure real-world applicability.[87]

   Benchmarking must shift from evaluating **token accuracy (ASR)** to measuring **semantic correctness and utility** within the **end-to-end conversational AI pipeline.** This shift enables evaluation of how effectively the entire system performs in real-world conditions, including its ability to handle grammar and sentence-structure errors in spoken input, as well as how errors propagate between different components of the system. Notably, model rankings can change significantly when evaluated based on the final LLM output's answer quality rather than intermediate transcription accuracy.[88]

   In addition, there is an urgent need for **standardised, usable testing frameworks** that evaluate the entire conversational pipeline under real-world conditions, including noise, latency, and interruption handling, to facilitate efficient government procurement and industry comparison.

86 Sujith Pulikodan, Prasanta Kumar Ghosh, Visruth Sanka, and Nihar Desai, "An Approach to Measuring the Performance of Automatic Speech Recognition (ASR) Models in the Context of Large Language Model (LLM) Powered Applications," arXiv preprint arXiv:2507.16456 (2025), https://arxiv.org

87 Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Mukund S. Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya, "Humanibench: A Human-Centric Framework for Large Multimodal Models Evaluation" arXiv pre print arXiv:2505.11454 (2025).

88 Inputs from workshop dated November 7, 2025.

f. **Managing accent and pronunciation variations:** To improve generalisation across accent and pronunciation variations, incorporate diverse speech datasets that capture regional differences. IndicVoices, for instance, has speakers from varied demographics across 408 districts.[89] Regular fine-tuning on region-specific data, along with adversarial training techniques to reduce accent bias, can further enhance generalisation. Additionally, voice cloning and accent correction techniques may be incorporated to tackle this challenge while building the solution.[90]

g. **Managing noisy speech in datasets:** Noisy speech data, common in Indian settings, can be addressed by evaluating signal quality like SNR and rejecting low-quality speech data. Noise-robust modeling techniques, such as synthetic noise augmentation and denoising autoencoders, can be employed to improve model performance in noisy environments.[91,92]

h. **Model cards:** Complementing datacards, model cards serve as structured documents that provide essential context and transparency for trained AI models. Mitchell et al. (2019)[93] suggest that model cards support the understanding of machine learning models, strengthen accountability, and promote responsible deployment.

Model Card Components

I. **Model details:**
   A. Basic information includes model name, version, developers, and release date.
   B. Training data sources: A description of the datasets used (e.g., Common Crawl, LibriSpeech, Wikipedia, or custom datasets), including whether the data is public or proprietary, as well as its domain and language coverage.

II. **Intended use:** Explicitly stating the scenarios and user groups for which the model is designed.

89 Javed et al., “IndicVoices.”
90 Hussam Azzuni and Abdulmotaleb El Saddik, “Voice Cloning: Comprehensive Survey,” arXiv preprint arXiv:2505.00579 (2025).
91 Karla Pizzi, Matías Pizarro, Asja Fischer, “Comparative Study on Noise-Augmented Training and its Effect on Adversarial Robustness in ASR Systems”, https://arxiv.org/
92 Đorđe T. Grozdić, Slobodan T. Jovičić, Miško Subotić “Whispered speech recognition using deep denoising autoen coder”, https://www.sciencedirect.com/
93 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, “Model Cards for Model Reporting,” in Proceedings of the Conference on Fairness, Accountability, and Transparency, (New York: ACM, 2019), 220-229.

III. **Performance metrics:** Reporting performance across relevant metrics (e.g., Word Error Rate for ASR, BLEU (Bilingual Evaluation Understudy) score for MT) and, crucially, performance breakdowns by demographic groups, accents, and languages. Indic voice models require reporting performance on various Indian accents, as done in Svarah and regional Hindi accents in LAHAJA.[94]

IV. **Limitations:** Detailing known failure modes, biases, and contexts where the model may perform poorly. This includes acknowledging limitations in low-resource Indic languages or specific dialects.

V. **Ethical considerations:** Discussions on potential societal impacts, fairness assessments, and measures taken to address biases. The ”Fair Tales” work on evaluating biases in Indian contexts serves as a prime example of the critical need for such assessments.[95]

i. **Model Deployment and Optimisation Strategies:** To address limited network coverage, optimised offline models can be deployed to ensure reliable operation without continuous connectivity. In addition, hybrid offline–online strategies allow systems to balance local processing with cloud capabilities when connectivity is available. To further reduce end-to-end latency, speech-to-speech models can be leveraged to minimise intermediate processing steps and enable faster, more natural interactions.

94 Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, and Mitesh M. Khapra, “LAHAJA: A Robust Multi-Accent Benchmark for Evaluating Hindi ASR Systems,” arXiv preprint arXiv:2408.11440 (2024). https://arxiv.org/.

95 Janki Atul Nawale, Mohammed Safi Ur Rahman Khan, Mansi Gupta, Danish Pruthi, and Mitesh M. Khapra, “Fairl Tales: Evaluation of Fairness in Indian Contexts with a Focus on Bias and Stereotypes,” arXiv preprint arXiv:2506.23111 (2025).

4. Embedding Responsible AI Practices

4. Embedding Responsible AI (RAI) Practices

The fragmented nature of the Indian voice technologies ecosystem has resulted in a lack of standardised approaches to data management, documentation, and ethical oversight, particularly in low-resource language contexts. This leads to reduced data quality, undermines interoperability, and obscures sources of bias. Gaps in information about contributors, labelling processes, and linguistic coverage ultimately make it difficult to identify whose voices are excluded, thereby amplifying the risk of encoding existing social inequalities into voice technologies.

In addition to the data management and sharing practices, developers must ensure compliance with legal obligations and responsible best practices from the data collection stage. Decisions made during the collection and curation process on intellectual property rights and data protection safeguards directly affect how the dataset can be used, shared, and repurposed across its lifecycle. In the section below, we outline key challenges and best practices for developers seeking to follow responsible AI (RAI) practices.

4.1 Challenges in Operationalising RAI Practices

a. **Inequitable data practices:**[96] At the data collection and curation stage, certain commonly used approaches—like crowdsourcing speech data without reciprocal agreements—can result in situations where providers of speech data, annotators and organisers do not share in the benefits acquired from these datasets. Such practices may also give rise to legal ambiguities surrounding the purposes for which such legacy speech data may be used, particularly in the absence of clear agreements.

b. **Ambiguous legal compliance mechanisms:** Under Indian law, particularly the Copyright Act, 1957, multiple layers of intellectual property protection may apply: to underlying text or transcripts (as literary works[97]), voice recordings (as sound recordings[98]), and curated metadata, provided each meets originality requirements. The use of copyrighted information requires a license from the person who owns the copyright. However, copyright does not extend to raw, unstructured data, such as random, uncurated recordings or isolated data points, due to a lack of originality, leaving such material free to be used without restriction.

Specific components of a speech dataset may qualify as personal information under Indian data protection law, including under the Digital Personal Data Protect Act, 2023 (DPDP Act). Where exemptions pertaining to use of publicly available data or use for research purposes do not apply,developers may be required to comply with data protection laws, including obtaining valid, informed consent from individuals and ensuring the lawful processing, storage, and transfer of such data, where applicable. The challenges of obtaining consent is further compounded when voice recordings are sourced from secondary sources such as previously published audio content, online archives, or public platforms that may contain personal data, given that developers are unlikely to have direct interface with the individuals concerned.

96 Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter et al., “The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages,” arXiv preprint arXiv:2502.15916 (2025), https://arxiv.org/
97 Section 2(o), (Indian) Copyright Act, 1957
98 Section 2(xx), (Indian) Copyright Act, 1957

4.2 Best Practices for Operationalising RAI Practices

To ensure the collection of voice datasets in a legally compliant and ethical manner, the following best practices may be considered:

a. **Meaningful engagement with language communities:** Data compilation process for speech and language technology must begin with meaningful engagement with associated language communities.[99] This involves understanding their needs and aspirations with respect to language technology and ensuring they have a say in how their data is used.[100] Such an approach is essential to avoid repeating a long-standing pattern of extractive practices, where communities, particularly those in the Global South, have provided information to academic institutions, governments, and multinational corporations without adequate value sharing.[101]

Participate in open voice technologies ecosystem

Ensure that datasets and tools are openly available to foster collaboration and accessibility, and promote further research based on your work. IndicVoices provides an open source blueprint for scalable data collection.[102] Karya's Attribution-NonCommercial-ShareAlike-FreeSoftware (BY-NC-SA-FS) 1.0 is one example of an open license that is designed to ensure that rights of contributors are preserved.[103] This license requires that credit is always given to the creator (BY), prevents commercial use (NC), mandates that adaptations are shared under the same terms (SA), and ensures that any incorporating software remains open under the GNU General Public License (FS). These elements uphold attribution and ensure that the benefits of community-contributed resources are fairly

99 Nina Markl, Lauren Hall-Lew, and Catherine Lai, "Language Technologies as if People Mattered: Centering Communities in Language Technology Development," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (Association for Computational Linguistics, 2024), 10085–10099.
100 Markl et al., "Language Technologies," 2024.
101 Bitange Ndemo, "Addressing Digital Colonialism: A Path to Equitable Data Governance," UNESCO Inclusive Policy Lab, 2024, https://es.unesco.org/inclusivepolicylab/analytics/addressing-digital-colonialism-path-equitable-data-governance?
102 Ndemo, "Addressing Digital Colonialism," 2024.
103 Supplied to authors by Karya.

shared and sustained within the open ecosystem. Having said that, there is a possibility that placing restrictions on reuse impacts the uptake of open licenses and their use.

b. **Documentation and packaging:** Comprehensive documentation and standardised packaging are essential for the usability, reproducibility, and long-term maintenance of voice datasets and models. This includes detailed guides on accessing, using, and contributing to the data and clear instructions for model deployment and fine-tuning.

   Adequate documentation should cover data schemas, annotation guidelines, and preprocessing steps, ensuring researchers and developers can correctly utilise the resources and contribute to the ecosystem. Packaging should follow established conventions to facilitate integration into various research and production pipelines. For example, researchers working on Project Vaani sliced the data by district, enabling efficient downloading and dataset management. Detailed documentation on HuggingFace[104] outlines all the metadata and data distribution, further supporting streamlined access and usage. In addition, lossless compression techniques were employed to reduce file size without compromising data quality.

   Maintaining comprehensive documentation practices should be viewed not merely as a compliance requirement, but as a safeguard against second-order harms such as downstream misuse and reputational damage. To this end, all consent (where applicable), rights clearances, licenses, and terms of use must be documented in a machine-readable format. Such documentation should specify the type and provenance of data used or created (including synthetic or derivative data), provide clear notice of liability where ethically or legally sensitive content may be involved, and flag any contextual limitations on use. It should also distinguish between general-purpose datasets and those developed for specific use cases (e.g., in health or legal domains), which may be subject to more restrictive terms. By adopting this approach, dataset creators can better mitigate risk, demonstrate accountability, and ensure responsible use across the data lifecycle.

c. **Ethical and legal pre-design:** The consent mechanism, PII reduction protocols, parameters for RAI as applicable to the specific use case, and data storage boundaries must be designed upfront rather than

104 Artpark, “Datasets,” https://huggingface.co/

treated as an afterthought, to ensure that the use of these datasets remains compliant with privacy laws and ownership boundaries.[105] Clear protocols and technical capabilities for human intervention should be established for situations in which AI systems encounter failures, exhibit harmful behaviours, or operate outside their intended parameters.[106] In the context of voice AI, this may involve human-in-the-loop validation for critical applications or mechanisms to flag and review potentially biased or incorrect outputs.

d. **Compliance with data protection laws:** Where possible, developers should obtain anonymised datasets that do not contain any PII. In cases where exemptions under the DPDP Act are not applicable, developers should ensure that clear and unambiguous consent in the form of an affirmative action is obtained from all participants, explaining how their data will be used, stored, and shared. For participants under the age of 18, ensure verifiable parental/guardian consent is obtained as prescribed. The DPDP Act requires granular consent and notice mechanisms that clearly specify distinct purposes of usage. Consent must extend to the processing of PII across different parts of the lifecycle, including: (1) initial audio recording, (2) transcription and annotation, (3) publicly hosting and distributing these datasets, and (4) potential voice synthesis applications. This consent request must be accompanied by a notice in clear and plain language with options to access content in English or any of the 22 languages specified in the Constitution's Eighth Schedule.

When voice recordings or transcripts are obtained from secondary sources, dataset creators may explore alternative legal bases for processing personal data, such as the exemptions for processing publicly available data where applicable, as the jurisprudence on the DPDP Act, evolves. Further, where feasible, developers should obtain contractual safeguards from secondary sources regarding the legality of the datasets provided.

Overall, developers should implement robust privacy protections, including ensuring compliance with data protection laws like the DPDP Act or General Data Protection Regulation (GDPR) for international collaborations. Privacy Enhancing Technologies (PETs)—for instance, analysing voice patterns on the fly without storing PII—should be actively explored and supported by the ecosystem.

105 Best practices shared by participants from a technical sprint workshop dated November 7, 2025.
106 Luciano Floridi and Josh Cowls, "A Unified Framework of Five Principles for AI in Society," Harvard Data Science Review 1, no. 1 (2019), https://doi.org/10.1162/99608f92.8cd550d1

e. **Compliance with copyright laws:** Under Indian law, particularly the Copyright Act, 1957, multiple layers of intellectual property protection may apply: to underlying text or transcripts (as literary works[107]), to voice recordings (as sound recordings[108]), and to curated metadata, provided each meets originality requirements. The use of any copyrighted information requires a license from the copyright holder. However, copyright does not extend to raw, unstructured data like random, uncurated recordings or isolated data points, due to lack of originality, leaving such material free to be used without restriction.

Where speech dataset collection involves primary contributions by individuals, dataset creators must secure appropriate rights through clearly documented assignments or licenses. These should be granted in exchange for commensurate consideration and must ensure that the developer has lawful authority to use, modify, and share the material. For voice datasets incorporating data from secondary sources (including but not limited to online content, archival materials, previously published works, or third-party datasets), special care must be taken to verify the original licensing terms and platform-specific terms of use, in order to avoid legal issues. To minimise complexity, it is recommended that creators rely, where feasible, on materials licensed under public domain dedications (e.g., CC0) or permissive open licenses (e.g., CC BY).

For each data point or resource incorporated from a secondary source, it is good practice to document the legal basis for its use, along with relevant metadata such as the source identifier (e.g., URL or publication reference), date of access or collection, applicable license terms, and any usage restrictions. The method used to verify permissions should also be recorded; this may include reviewing licensing statements, platform terms of use, machine-readable license metadata, or technical signals (e.g., robots.txt, content headers, or repository tags). Adhering to such documentation practices helps support compliance with copyright and licensing obligations, as well as emerging data governance standards.[109]

f. **Ensure transparency:** It is crucial that developers are transparent about data collection and management practices with participants and stakeholders. Users must be provided with clear information about how voice AI systems work, as well as their capabilities and

107 Section 2(o), (Indian) Copyright Act, 1957
108 Section 2(xx), (Indian) Copyright Act, 1957
109 Stefan Baack et al., “Towards Best Practices for Open Datasets for LLM Training,” arXiv preprint arXiv:2501.08365 (January 14, 2025),https://arxiv.org/.

limitations.[110] Distinguishing between human and synthetic voices can help mitigate risks of fraud and misinformation.[111]

g. **Foster accountability:** Mechanisms for accountability, including regular audits and clear governance structures, must be established, and a responsible individual or team should be designated for data protection and compliance. Continuous monitoring of AI system performance in real-world deployment is essential to detect anomalies, performance degradation, and emergent issues.[112] This involves tracking error rates, latency, and resource utilisation across diverse user demographics and linguistic variations.[113]

In addition, developers should continuously evaluate deployed models for biases and stereotypes—particularly those related to caste, religion, region, and tribal identities, which are critical in the Indian context. This requires understanding how pre-existing societal prejudices may be reflected or amplified in model outputs, and developing strategies to counter them, such as debiasing techniques or adding post-processing filters.[114] Developers should aim to ensure that voice AI systems provide equitable access and performance for all users, regardless of linguistic background, accent, or socio-economic status.[115]

h. **Identifying the license under which open-sourcing is carried out:** Licensing datasets and models is a critical decision that dictates permissible scope of use, modification, and distribution. Commonly used licenses for open-sourcing efforts in the Indic language domain include Creative Commons (CC) licenses. For instance, the Bharat Parallel Corpus Collection (BPCC) and associated datasets from the IndicTrans2 project are released under permissive licenses such as CC0 and CC-BY-4.0. CC0 dedicates the work to the public domain, allowing maximum freedom of use, while CC BY 4.0 requires attribution. IndicVoices speech datasets were also released under the permissive CC-BY-4.0 license, with the tools being released under an MIT license[116], ensuring broad access for academic, research, and commercial applications. Selecting appropriate

110 Kate Crawford, The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. (New Haven: Yale University Press, 2021).
111 Nicholas Carlini and David Wagner, “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text,” in 2018 IEEE Security and Privacy Workshops (SPW) (IEEE, 2018), 1–7,https://doi.org/
112 Luciano Floridi, and Josh Cowls, “A Unified Framework of Five Principles for AI in Society,” Harvard Data Science Review 1, no. 1 (2019). https://doi.org/
113 Timnit Gebru, et al., “Datasheets for Datasets,” Communications of the ACM 64, no. 4 (2021): 86–92. https://doi.org/
114 Nithya Sambasivan et al., “‘Everyone Wants to Do the Model Work, Not the Data Work’: Data Cascades in High-Stakes AI,” In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, (ACM, 2021), 1–15, https://doi.org/
115 Crawford,The Atlas of AI, 2021.
116 Nivash Jeevanandam, “IIT Madras, AI4Bharat, and Sarvam AI Launch IndicVoices: A Milestone in Indian Speech Recognition,” IndiaAI, November 28, 2024, https://indiaai.gov.in/.

licenses provides legal clarity for downstream users, promotes widespread adoption, and facilitates collaborative development within the community. Organisations may also consider a tiered licensing approach—for instance, releasing the audio components under a CC BY 4.0 license, and the texts/transcripts under different licenses—to encourage open derivatives while separating text copyright from voice rights. While permissive licenses such as CC0 may often be a default choice, organisations should evaluate the context and objectives of the project before finalising licensing decisions.

Licensing Considerations for Developers

- **Provenance of the data:** If a dataset incorporates or builds upon third-party data, its original license may impose obligations (e.g., ShareAlike requirements under CC-BY-SA license) that can taint the entire dataset and limit downstream flexibility. Downstream developers must be mindful of the obligations they have when using existing datasets.

- **Desired openness:** If encouraging broad reuse and integration into commercial workflows is a priority, more permissive licenses may be appropriate. However, if retaining openness in derivative works is critical, a CC-BY-SA license may be preferable.

- **Component-specific considerations:** Licensing may differ across modalities — for example, audio files may be licensed separately from textual transcripts. A tiered approach across licenses can help manage distinct rights while promoting reuse and protecting derivative contributions.

- **Commercial considerations:** Licensing choices will also vary depending on whether the intended end use is commercial or non-commercial.

- **Encouraging innovation and re-use:** Where the primary goal is to promote diverse and inclusive development of voice technology solutions, choosing a permissive license can support this goal with due consideration of privacy, fairness and data protection.

Organisations may also consider exploring the Creative Commons license selector tool[117], which builds upon the abovementioned factors.

i. **Terms of use:** Developers should adopt Terms of Use and Acceptable Use Policies that define how data may be accessed, shared, modified, and redistributed. These policies should explicitly address data protection considerations, attribution, commercial use, and downstream redistribution, in order to ensure legal compliance and ethical usage. For example, Mozilla's Common Voice dataset uses the Creative Commons CC0 license, allowing unrestricted use but clearly stating in its terms[118] that data is contributed with informed consent and should not be used in ways that violate privacy, promote harm, or enable the identification of individuals. Furthermore, similar to model cards, these policies should clarify that different datasets may be subject to other terms, including varying degrees of permissive use tailored to specific use cases. Adopting well-defined and transparent usage terms helps protect both data contributors and consumers while promoting trust and alignment with responsible AI development practices. Such policies should also make clear that the user bears responsibility for downstream use. Incorporating Responsible AI (RAI) license terms can further enhance clarity around what constitutes harmful or unlawful downstream use of the technology.

   RAI licenses allow stipulating the terms of use of an AI system or an AI model and impose restrictions to prevent harmful applications. These licenses typically include the terms of the licenses, standard limitations of liability, and end-use related prohibitions. For instance, a license may require compliance with conduct rules that prohibit unlawful, harmful, misleading, discriminatory, exploitative, or high-risk uses, with violations subject to suspension or termination of access.

j. **Selecting a hosting platform:** The selection of a hosting platform for open-source datasets and models should be guided by considerations of accessibility, scalability, version control, and community engagement features. Commonly used platforms include GitHub, Hugging Face, and dedicated institutional repositories.[119] Additional factors to consider include data sovereignty; platforms such as IndiaAI's AI-Kosha (a centralised repository of datasets and

117 Creative Commons, "Choose a License for Your Work," https://creativecommons.org//
118 Mozilla Foundation, "Common Voice Terms of Use," https://commonvoice.mozilla.org/.
119 Please see, Hugging Face, https://huggingface.co/ and GitHub, https://github.com/, accessed July 2025.

models)[120], and Bhashini's ULCA (standard API and open scalable data platform supporting various types of datasets for Indian languages datasets and models),[121] offer indigenous, government-managed platforms.

Hosting Platforms for Open Data and Models: Key Considerations

- **Accessibility:** Ensuring broad access for researchers and developers worldwide, particularly those in resource-constrained environments
- **Scalability:** The ability to host large volumes of data and models, accommodating future growth
- **Version control:** Robust versioning systems to track changes, updates, and different iterations of datasets and models
- **Community engagement:** Features that enable collaboration, feedback, and contributions from the more exhaustive research and developer community
- **Security and reliability:** Measures to protect data integrity and ensure consistent availability
- **Sustainability:** Adequate financial and governance capacity to support long-term hosting of data and models

k. **Audit Techniques and Benchmarks:** Audit techniques and robust benchmarks should be used to evaluate system performance against specified safety and fairness criteria throughout the deployment lifecycle.[122] For Indic languages, this would involve ongoing evaluation against diverse accent and dialect benchmarks to ensure equitable performance.[123, 124]

From an RAI perspective, the downstream use of voice technology in India should account for broader societal implications and ensure fairness and equity, especially given the country's unique socio-cultural dynamics. To conduct these assessments, developers can

120 AIKOSH IndiaAI, https://aikosh.indiaai.gov.in/home
121 Bhashini ULCA, https://bhashini.gov.in/ulca
122 Floridi and Cowls, "Unified Framework of Five Principles for AI in Society."
123 Sambasivan et al., "'Everyone Wants to Do the Model Work, Not the Data Work,'" CHI 2021.
124 Gebru et al., "Datasheets for Datasets."

draw on existing frameworks. Responsible AI Assessments is one such framework, designed to support developers in conducting holistic AI risk and ethics assessments by bridging the gap between principles and practical implementation.[125] It also offers suitable mitigation strategies for developers to apply to their specific contexts.

l. **Mitigating misuse:** AI systems should undergo regular and rigorous assessment of their safety properties, including robustness to adversarial attacks and resilience under unforeseen circumstances.[126] This is particularly important for voice systems susceptible to audio manipulations or voice impersonation attacks.[127] Appropriate mitigation mechanisms—such as avoiding the collection of PII, ensuring appropriate licensing, and using watermarks—should be considered for open datasets and models.

125 GIZ FAIR Forward and Eticas, "Responsible AI Assessments: Identify and Assess Potential Harms and Biases in AI Systems with a Focus on Use Cases in Sub-Saharan Africa and Asia," 2024,https://openforgood.info/
126 Floridi and Cowls, "Unified Framework of Five Principles for AI in Society."
127 Nicholas Carlini and David Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," In 2018 IEEE Security and Privacy Workshops (SPW) (IEEE, 2018),1–7, https://doi.org/.

Appendix 1

Current Status in Open Source Voice Technologies in India

Appendix 1: Current Status in Open Source Voice Technologies in India

The voice-tech ecosystem in India has made tremendous progress, driven by various government initiatives, university-led efforts, open-source contributions, and the broader democratisation of technology by multiple organisations. This growth is also supported by rapid advancements in machine learning worldwide.

The ecosystem is built around datasets, models, frameworks and tools that collectively enable and accelerate voice technologies. These components form the essential foundation for building and scaling impactful solutions. While some elements are universally applicable, others need to be adapted to the unique linguistic and cultural requirements of the Indian context. Each component of the ecosystem is at a different stage of maturity, yet all are vital for developing and deploying solutions capable of serving billions of people.

The Indian government has strengthened the AI ecosystem through initiatives such as BHASHINI[128] and the IndiaAI Mission[129]. BHASHINI bridges language, literacy, and digital divides by offering voice-first platforms, models, and datasets. BHASHINI helps to make government services accessible across India, especially in rural areas, in collaboration with ministries, startups, and private organisations. The IndiaAI Mission supports AI innovation by providing computing access and promoting ethical, inclusive, and indigenous AI development through platforms like AI Kosh,[130] which offers datasets, models, and tools to enable AI innovation.

1. Datasets

Building effective machine learning models requires high-quality, comprehensive datasets, as the accuracy of voice AI systems depends heavily on the strength of the underlying corpora. The dataset creation process involves data recording, data cleaning, annotation and

128 Bhashini, https://bhashini.gov.in/
129 IndiaAI, https://indiaai.gov.in/
130 AIKosh, https://aikosh.indiaai.gov.in/home

transcription, segmentation, normalisation, quality assurance, and language-specific curation. India, known for its exceptional linguistic diversity—with 22 officially recognised languages and more than 700 living languages—presents a unique challenge[131]. Even within a single language, significant regional variations exist. Creating a corpus that captures this level of diversity demands extensive effort and careful design.

There are several initiatives aimed at creating datasets for building voice technologies, including Vaani[132], IndicVoices[133], Kathbath[134], SPRING-INX[135], GramVaani[136], RESPIN[137], and SYSPIN[138]. A large number of these initiatives are supported by Bhashini.

Augmenting these foundational datasets with use-case-specific data can improve model performance. Although this dataset includes multiple languages and tens of thousands of hours of audio data, it is still much smaller than Western resources such as VoxPopuli, a large-scale multilingual corpus offering around 400,000 hours of unlabelled speech across 23 languages.[139]

DATASET	SELECTION METHOD	TOTAL DURATION (HRS)	#LANG.	#SPEAKERS
IndicVoices	read (8%), extempore (76%), conversational (16%)	19550	22	29K
Shrutilipi	read & conversational	6457	12	—
Kathbath	read	1684	12	1218
Spring-inx	read, extempore & conversational	2005	10	7609
FLEURS	read	163	13	—

131 Government of India, “Census 2011, Language”, 2018.
132 Artpark, “VAANI: Capturing the Language Landscape for an Inclusive Digital India.”
133 Tahir et al., “IndicVoices.”
134 AI4Bharat, “ IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian languages.”
135 Spring Lab, IIT Madras, “Spring-inx: A Multilingual Indian Language Speech Corpus.”
136 “Gram Vaani ASR Challenge Dataset,” https://sites.google.com/view/gramvaaniasrchallenge/dataset
137 Kumar et. al., “RESPIN-S1.0.”
138 “SYSPIN_S1.0 Corpus: A TTS Corpus of 900+ Hours in Nine Indian Languages“
139 Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, “VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Association for Computational Linguistics, 2021), 993–1003, https://doi.org/

Common Voice	read	373	10	—
MUCS Dataset	read	351	6	158
IndicTTS	read	225	13	25
SYSPIN	read	900	9	18
RESPIN	read	10,000	9	18.8k
VAANI	spontaneous	31,255	109	143K

2. Models

Machine learning models form an integral part of the voice-tech ecosystem. Various types of models are used across different stages of voice-enabled solutions. At the core of this ecosystem are two primary model categories: Automatic Speech Recognition (ASR) and Text-to-Speech (TTS), which together form the foundation of modern voice technology.

ASR models

Many of the open-source ASR models available today for Indic languages are built upon architectures such as Whisper[140], Wav2Vec[141], Wav2Vec2[142], and Conformer[143]. These models generally follow a pre-training and fine-tuning paradigm, where the base model is trained on large multilingual corpora and later adapted to specific languages or datasets. For Indic languages, several models have undergone further pre-training with additional Indic speech, while many others have been fine-tuned directly using transcribed Indic datasets.

The Indic Conformer[144] model suite from AI4Bharat consists of fine-tuned Conformer encoders paired with either CTC or RNN-T decoders.

The collection includes 23 models: 22 monolingual models (120M parameters), each corresponding to one of India's official languages, and a larger 600M-parameter multilingual model capable of transcribing all 22 languages. The Vakyansh Toolkit[145] provides Wav2Vec2-based

140 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint, December 2022, https://arxiv.org/abs/2212.04356.
141 Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," arXiv preprint, April 2019, https://arxiv.org/
142 Alexei Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Advances in Neural Information Processing Systems 33 (NeurIPS, 2020), 12449–12460
143 Anmol Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," arXiv preprint, May 2020, https://arxiv.org/
144 AI4Bharat, "IndicConformer," Hugging Face Model Collection, https://huggingface.co/
145 Harveen Singh Chadha, et al., "Vakyansh: ASR Toolkit for Low Resource Indic Languages," arXiv preprint, March 2022, https://arxiv.org/

models pre-trained on 10,000 hours of multilingual audio spanning 23 languages and fine-tuned individually for languages such as Hindi,

Rajasthani, Maithili, Bhojpuri, Malayalam, Kannada, Telugu, Marathi, Bengali, Urdu, Tamil, and Odia. Vaani Whisper[146] models extend the Whisper architecture by fine-tuning it on Vaani data combined with other publicly available datasets. These models are offered in multiple configurations: Whisper-Small (Kannada, Hindi, Tulu), Whisper-Medium (Kannada, Hindi, Telugu, Bengali), and Whisper-Large (Hindi). pingala-v1-universal[147] from ShunyaLabs, Omnilingual ASR[148] from Meta, and Voxtral[149] from Mistral represents the latest generation of multilingual and omnilingual speech-recognition models, offering broad coverage across numerous Indian languages.

At the same time, we are seeing multimodal LLMs integrate speech understanding directly into their architectures, reducing or even eliminating the need for a separate ASR component in voice applications. Open-source models like Gemma 3N[150] are also advancing this trend by providing strong multilingual support, including many Indic languages.

TTS models

There are many open-source TTS models available today, supported by large-scale targeted data collection efforts such as SYSPIN, IndicTTS, and Indic Voices-R. Some notable models in this ecosystem include Indic F5[151], IndicTTS[152], Indic Parlor TTS[153], Veena[154], Chatterbox TTS[155], and CoquiTTS[156]. These models provide a wide range of capabilities, including voice cloning, expressive and emotional speech synthesis, and increasingly natural-sounding audio output.

3. **Tools and Frameworks**

Different tools are employed at various stages of the voice technology lifecycle to improve data collection, curation and downstream development of speech technologies. While some tools are broadly applicable across general ML workflows, others are specifically designed for speech-related tasks. In the following sections, we will walk through each phase of the technology and highlight some of the most widely used tools.

146 ARTPARK-IISc, “Vaani Whisper,” Hugging Face Model Collection, https://huggingface.co/
147 Shunya Labs, “Pingala-v1-universal,” Hugging Face, https://huggingface.co/
148 Gil Keren et al. ,”Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages,” 2025, https://arxiv.org
149 Alexander H. Liu et al., “Voxtral,” arXiv preprint, 2025, https://arxiv.org/
150 Google DeepMind, “Gemma 3n,” https://deepmind.google/
151 AI4Bharat, IndicF5, GitHub, https://github.com/
152 AI4Bharat, Indic-TTS, GitHub, https://github.com/
153 AI4Bharat, “indic-parler-tts,” https://huggingface.co/
154 Maya Research, “Veena,” Hugging Face, https://huggingface.co/
155 Resemble AI, chatterbox, GitHub, https://github.com/resemble-ai/
156 Coqui AI, TTS, GitHub, https://github.com/coqui-ai/

Speech data collection

Speech data collection involves obtaining data either from primary sources or by leveraging secondary sources such as the internet. There are mainly three types of speech data, such as read speech, spontaneous speech, and conversational speech[157]. When collecting data from primary sources, a recording application, typically mobile or web-based application is required. The recording application must be designed to support the specific type(s) relevant to the intended use case. A variety of tools, both open-source and commercial, are available to facilitate this process.

TOOL NAME	PUBLISHER	PLATFORM	LICENSE
Lingua Recorder[158]	Lingua-libre	Cross-browser voice recording JS library	MIT
Karya[159]	Karya Inc.	Web, Android	GPL-3
Kathbath[160]	AI4Bharat	Web, Android, IOS	MIT
LiG_AIKUMA[161]	Université Grenoble Alpes	Android	AGPL
Open Data Kit (ODK)[162]	GetODK Inc	Android	Apache 2.0

Speech labelling

The labelling process depends on the type of model being developed. For ASR, transcripts must be generated for each audio segment. For Language Identification (LID), it is sufficient to assign a language label to each segment. Likewise, for other classification tasks, only the corresponding class label is needed. In the case of segmentation tasks, the start and end points of each segment must be annotated.

Several tools are available for audio editing and annotation, each is suited to different needs.

Audacity is a general-purpose, open-source audio editor with basic annotation features, widely used for simple tasks. Praat is a favorite in linguistics, offering advanced phonetic analysis and annotation capabilities. ELAN supports complex, multi-layer time-aligned

157 Philipp Gabler, Bernhard C. Geiger, Barbara Schuppler, and Roman Kern, “Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition,” Information 14, no. 2 (2023): 137, https://doi.org/.
158 Lingua Libre, LinguaRecorder, GitHub, https://github.com/
159 Karya Inc., GitHub, https://github.com/
160 AI4Bharat, “Kathbath,” https://ai4bharat.iitm.ac.in/
161 Lig-Aikuma, https://lig-aikuma.imag.fr/
162 ODK, GitHub, https://github.com/getodk

annotations and is commonly used in linguistic and behavioural research. For managing and annotating speech databases via the web, EMU-SDMS provides a powerful interface. Sonic Visualiser specialise in detailed visualisations of audio such as spectrograms and waveforms, aiding in precise annotation. Label Studio is a versatile, open-source tool for multi-modal data annotation, including audio. Audino is a web-based annotation platform tailored to speech datasets, ideal for projects in machine learning and speech processing. Shoonya is an open-source platform designed for large-scale data annotation and labelling, with the goal of strengthening the digital presence of underrepresented Indian languages.

NAME	USE CASE	PUBLISHER	MODE	LICENSE
Audacity[163]	General purpose audio editing with basic annotation	Audacity Team	Desktop App.	GPLv3
Praat[164]	Phonetic analysis and annotation	University of Amsterdam	Desktop App.	GPLv3
ELAN[165]	Multi-layer time-aligned annotations	MPI	Desktop App.	GPLv3
EMU-SDMS[166]	Phonetic annotation and structured speech storage	IPS LMU	Web-based	MIT
Sonic Visualiser[167]	Visualisation and annotation (wave-form, spectrogram)	Centre for Digital Mu-Desktop sic, QMUL	Desktop App.	GPLv2
Label Studio[168]	Classification, speaker diarisation, emotion recognition, audio transcription	Heartex	Web-based	Apache 2.0
Audino[169]	Speech annotation tool for ASR datasets	MIDAS Lab	Web-based	MIT

163 Audacity, https://www.audacityteam.org/
164 Paul Boersma and David Weenink, “Praat: Doing Phonetics by Computer,” https://www.fon.hum.uva.nl/
165 Max Planck Institute, “ELAN: Annotation Tool for Audio and Video Recordings,” https://archive.mpi.nl/
166 IPS-LMU, “EMU Speech Database Management System (EMU-SDMS),” https://ips-lmu.github.io/
167 Chris Cannam, Christian Landone, and Mark Sandler, “Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files,” in Proceedings of the ACM Multimedia 2010 International Conference (New York: ACM, 2010). https://www.sonicvisualiser.org/
168 HumanSignal, “Label Studio: Open Source Data Labeling,” https://labelstud.io/
169 MIDAS Research, audino, GitHub, https://github.com/midas-research/audino

Shoonya[170]	Web based multilingual speech/text/image annotation for NLP datasets	AI4Bharat	Web-based	MIT

Speech data management

Several tools and platforms support efficient management, processing, and annotation of speech datasets across various research and development contexts. Lhotse is a Python library designed for organising and processing speech data, with compatibility for popular ASR frameworks like Kaldi. HuggingFace Datasets[171] provides a standardised interface for loading, sharing, and versioning datasets, including those for speech. Audino is a web-based tool tailored for annotating and managing audio, especially during the creation of ASR datasets. EMU-SDMS offers structured speech database management with a focus on phonetic annotation.

NAME	PUBLISHER	PLATFORM	LICENSE
Lhotse[172]	Johns Hopkins (CLI/API) University/ESP-net contributors	Python	Apache 2.0
Audino	MIDAS Lab, IIIT-H	Web-based	MIT
EMU-SDMS[173]	LMU Munich	Web-based	GPL

4. Model Development

Selecting the appropriate model architecture is a critical step in developing systems for ASR, TTS, or Speaker Identification (Speaker ID). In some cases, we fine-tune popular foundation models for downstream tasks using our own data, while in other cases, we train models from scratch. The latest architectures are primarily based on transformer-based designs. For ASR tasks, popular foundational models include Whisper, Wav2Vec 2.0, Conformer, Universal Speech Model (USM)[174], Massively Multilingual Speech (MMS)[175] and Fast conformer[176], each known for high accuracy and adaptability to various languages and

170 AI4Bharat, “Shoonya,” https://ai4bharat.iitm.ac.in/
171 Hugging Face, datasets, GitHub, https://github.com/
172 Lhotse, lhotse, GitHub, https://github.com/
173 IPS-LMU, “EMU Speech Database Management System (EMU-SDMS),” https://ips-lmu.github.io/
174 Yu Zhang, et al., “Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages,” arXiv preprint, March 2023.
175 Vineel Pratap et al., “Scaling Speech Technology to 1,000+ Languages,” arXiv preprint arXiv:2305.13516 (2023), https://arxiv.org/
176 Dima Rekesh, et al., “Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition,” arXiv preprint, May 2023, https://arxiv.org/

acoustic conditions. In TTS applications, models such as Tacotron[177], FastSpeech[178], and VITS[179] are commonly used, offering high-quality, natural-sounding synthetic speech. For Speaker ID, robust architectures like x-vector[180] and ECAPA-TDNN[181] are favored for their ability to capture speaker characteristics effectively.

Recently, significant progress has been made in multimodal models that incorporate speech as one of the modalities. Models such as Phi-4[182], Qwen2.5-Omni[183], and Voxtral, Gemma3N effectively integrates speech with other modalities. However, coverage of Indic languages in these models remains very limited. Additional fine-tuning can help extend their capabilities to support a broader range of Indic languages.

These models are typically implemented using frameworks such as PyTorch[184], TensorFlow[185] and framework such as Hugging Face Transformers[186], ESPnet[187], Fairseq[188] and NeMo[189] which build on top provide extensive support for training, fine-tuning, and deployment of speech models in both research and production environments.

NAME	PUBLISHER	TYPE	LICENSE
Whisper	OpenAI	Automatic Speech Recognition (ASR)	MIT
Wav2Vec 2.0	Facebook AI Research (FAIR)	Self-supervised ASR model	MIT
Conformer	Google Research	Hybrid CNN-Trans-former ASR model	—
Tacotron	Google	Text-to-Speech (TTS)	—
FastSpeech	Microsoft	Fast TTS synthesis model	—

177 Yuxuan Wang et al., “Tacotron: Towards End-to-End Speech Synthesis,” arXiv preprint arXiv:1703.10135 (2017), https://arxiv.org/

178 Yi Ren et al., “FastSpeech: Fast, Robust and Controllable Text to Speech,” arXiv preprint arXiv:1905.09263 (2019), https://arxiv.org/

179 Jaehyeon Kim, Jungil Kong, and Juhee Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” arXiv preprint, June 2021, https://arxiv.org/

180 Daniel Povey et al., “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2018), https://www.danielpovey.com/

181 Brecht Desplanques et al., “ECAPA-TDNN: Emphasised Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” arXiv preprint arXiv:2005.07143 (2020), https://arxiv.org/

182 Marah Abdin, et al., “Phi-4 Technical Report,” arXiv preprint, December 2024,https://arxiv.org/

183 Jin Xu, et al., “Qwen2.5-Omni Technical Report,” arXiv preprint, March 2025, https://arxiv.org/

184 Adam Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” arXiv preprint arXiv:1912.01703 (2019), https://arxiv.org/

185 Martín Abadi et al., “TensorFlow: A System for Large-Scale Machine Learning,” in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (2016), 265–283,https://arxiv.org/

186 HuggingFace, “Transformers: State-of-the-art Natural Language Processing,” https://huggingface.co/.

187 Shinji Watanabe et al., “ESPnet: End-to-End Speech Processing Toolkit,” arXiv preprint arXiv:1804.00015 (2018); “ESPnet2: End-to-End Speech Processing Toolkit for 2000+ Languages,” arXiv preprint arXiv:1910.03771 (2019), https://arxiv.org/

188 fairseq: A Fast, Extensible Toolkit for Sequence Modeling

189 NVIDIA, NeMo, GitHub, https://github.com/

VITS	NAVER AI Lab	End-to-end TTS with GANs	MIT
ECA-PA-TDNN	Ghent University	Speaker embedding/ verification	—
Universal Speech Model (USM)	Google	Automatic Speech Recognition (ASR)	—

These foundational models offer well-tested architectures that effectively capture speech characteristics, enabling easy adaptation to end-use applications with minimal data. These models are typically trained on vast amounts of data; however, in the Indian context, challenges remain due to limited coverage of many languages and comparatively less data available for mainstream Indian languages than for their Western counterparts.[190]

Model Training

Model training for speech tasks such as ASR, TTS, or Speaker ID typically requires the use of GPUs to handle the computational demands of deep learning architectures. Efficient training involves not only optimising model performance but also maintaining visibility into the training process. This is achieved through robust logging and monitoring tools. BitsAndBytes[191] facilitates the use of LLMs in PyTorch through k-bit quantisation, offering three key features that significantly reduce memory consumption during both training and inference. PyTorch Lightning[192]simplifies training loops and supports scalable, modular code, while Weights & Biases[193] and TensorBoard[194] provide powerful visualisation dashboards for tracking metrics, loss curves, model checkpoints, and system usage. Together, these tools enable reproducible experiments, early detection of training issues, and better model performance through informed decision-making.

190 Radford et al., “Robust Speech Recognition.”
191 bitsandbytes-foundation, bitsandbytes, GitHub, https://github.com/
192 Lightning AI, pytorch-lightning, GitHub, https://github.com/
193 Weights & Biases, wandb, GitHub, https://github.com/
194 TensorFlow, “TensorBoard,” https://www.tensorflow.org/

NAME	PUBLISHER	AREA	LICENSE
PyTorch	Meta	Deep learning framework	BSD
TensorFlow	Google	Deep learning framework	Apache 2.0
Hugging Face Trans-formers	Hugging Face	Pretrained models for NLP, speech, vision	Apache 2.0
Weights & Biases	Weights & Biases Inc.	Experiment tracking and model monitoring	MIT
ESPnet	Waseda University / Con-tributors	End-to-end speech processing toolkit	Apache 2.0
fairseq	Meta AI	Sequence modeling toolkit (ASR, MT, TTS)	MIT
Speech-Brain	SpeechBrain Team / JHU	Open-source speech toolkit (ASR, TTS, speaker ID)	Apache 2.0
Accelerate	Hugging Face	Multi-GPU, TPU, mixed-precision training wrapper	Apache 2.0
DeepSpeed	Microsoft	Large-scale distributed training	MIT
NeMo	Nvidia	Scalable generative AI framework	Apache 2.0

Model Evaluation

NAME	AREA	METRIC	LICENSE
Jiwer[195]	ASR	WER, CER	GPL
Hugging Face Evaluate[196]	ASR	WER, CER	Apache 2.0
MOSNet[197]	TTS	MOS Prediction	MIT
pysepm[198]	Speech Quality	PESQ, STOI, SDR, SNR	GPL
pyannote-audio[199]	Speaker Diarisation	Embedding Similarity, Diarisation Error Rate	MIT
speechmetrics[200]	General Speech Evaluation	WER, MOS, STOI, PESQ, etc.	Apache 2.0

195 Jitsi, “jiwer,” https://jitsi.github.io/jiwer/
196 Hugging Face, evaluate, GitHub, https://github.com/
197 Chen-Chou Lo, et al., “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” in Interspeech 2019 (ISCA, 2019), https://arxiv.org/abs/
198 Philipp Schmid, pysepm, GitHub, https://github.com/
199 pyannote.ai, pyannote-audio, GitHub, https://github.com/
200 Ali Utku, speechmetrics, GitHub, https://github.com/

Evaluating speech models involves using task-specific metrics to assess performance accurately. For ASR, metrics like Word Error Rate (WER) and Character Error Rate (CER) are commonly used to measure transcription accuracy. In Text-to-Speech (TTS) systems, Mean Opinion Score (MOS) is often employed to evaluate the naturalness and intelligibility of generated speech, typically through human listening tests or predictive models. Based on evaluation results, models may require fine-tuning on specific datasets or domains to improve performance, particularly when dealing with diverse accents, background noise, or underrepresented languages. This iterative evaluation and refinement process ensures that the model generalises well and meets the desired quality standards. Tools like MLFlow[201] provides a comprehensive platform for experiment tracking, enabling enhanced observability and streamlined evaluation of machine learning models.

NAME	PUBLISHER	MODE	LICENSE
TorchServe	Meta AI	Model serving for PyTorch (REST/ gRPC)	Apache 2.0
Triton Inference Server	NVIDIA	Multi-framework model serving (ONNX, TensorRT, PyTorch, TF)	BSD
FASTAPI	Sebastián Ramírez / Community	High-performance API framework (ASGI)	MIT
Flask	Pallets Projects	Lightweight WSGI web framework for APIs	BSD
TensorFlow Serving	Google	TensorFlow model serving via REST/ gRPC	Apache 2.0
ONNX	onnx.ai	Open standard for machine learn- ing interoperability	Apache 2.0
Kubeflow[202]	kubeflow.org	Addresses multiple aspects of the AI lifecycle	Apache 2.0
Feast[203]	feast.dev	Feature management for training and real-time inference	Apache 2.0

201 MLflow, https://mlflow.org/
202 Kubeflow, https://www.kubeflow.org/
203 Feast, https://feast.dev/

Several evaluation frameworks have been proposed for voice technologies in Indic languages. Some are model-specific, while others assess end-to-end systems. Notable examples include Vistaar[204], IndicSUPERB[205], and LAHAJA[206].

5. Deployment

Once a speech model is trained and evaluated, the next step is deployment and serving. To ensure efficient inference in production environments, models are often converted into optimised formats such as ONNX[207] or TorchScript[208], which enhance portability and performance across different platforms. For real-time or batch inference, these models can be served via REST or gRPC APIs, enabling seamless integration with downstream applications. Tools like Triton Inference Server[209] provide scalable, high-performance model serving with support for batching, version control, and multi-framework deployment. Alternatively, lightweight frameworks like FastAPI[210] offer flexibility for deploying custom APIs with minimal latency, making them suitable for smaller-scale or experimental deployments.

NAME	PUBLISHER	MODE	LICENSE
Prometheus[211] + Grafana[212]	CNCF / Grafana Labs	Monitoring, Visualisation	Apache 2.0 / AG-PLv3
Seldon Core[213]	Seldon	Model deployment, drift detection, monitoring	Apache 2.0
Airflow[214]	Apache Software Foundation	Workflow orchestration	Apache 2.0
MLFlow[215]	mlflow.org	Experiment tracking, observability, evaluation	Apache 2.0
Metaflow[216]	Netflix	Build and manage real-life AI/ML systems	Apache 2.0
Evidently AI[217]	Evidently AI	Evaluation, testing, monitoring for ML and LLM systems	Apache 2.0

204 AI4Bharat, vistaar, GitHub, https://github.com/
205 AI4Bharat, IndicSUPERB, GitHub, https://github.com/
206 AI4Bharat, Lahaja, GitHub, https://github.com/
207 ONNX, https://onnx.ai/
208 PyTorch JIT Documentation, https://docs.pytorch.org/
209 NVIDIA, “Dynamo Triton,” https://www.nvidia.com/
210 FastAPI, https://fastapi.tiangolo.com/
211 Prometheus, https://prometheus.io/
212 Grafana Labs, “Grafana,” https://grafana.com/
213 Seldon, seldon-core, GitHub, https://github.com/
214 Apache Software Foundation, “Apache Airflow,” https://airflow.apache.org/
215 MLflow, https://mlflow.org/
216 Metaflow, https://metaflow.org/
217 Evidently AI, evidently, GitHub, https://github.com/evidentlyai/

NannyML[218]	NannyML	Estimate post-deployment model performance	Apache 2.0
Alibi Detect[219]	Seldon	Outlier, adversarial, and drift detection	Apache 2.0

218 NannyML, nannyml, GitHub, https://github.com/
219 Seldon, alibi-detect, GitHub, https://github.com/

Appendix 2

Workshop Participants

Appendix 2: Workshop Participants

We gratefully acknowledge the following individuals for their active participation in the workshop sprints. Their discussions, insights, and diverse perspectives significantly informed the review and refinement of this toolkit.

WORKSHOP: OCTOBER 9, 2025 WITH BHASHINI
Responsible and Open Voice Technologies in India

1. Amitabh Nag
2. Shailendra Pal Singh
3. Karan Joshi
4. Trisha Singh

WORKSHOP: OCTOBER 16, 2025
Legal and Ethical Considerations in Indic Voice Tech Governance

1. Janki Nawale (AI4Bharath)
2. Vibhav Mithal (Anand and Anand)
3. Nikhil Narendran (Trilegal)
4. Aparajita Lath (NLSIU)
5. Swaraj Barooah (SpicyIP)
6. Shreya Ramann (Counselect)

WORKSHOP: NOVEMBER 7, 2025
Data and Modelling for Open-Source Indic Voice Tech

1. Anurag Behera (Srujnee)
2. Chintan Parikh (Reverie)
3. Chockalingam Muthian (NASSCOM)
4. Santosh Kevlani (EkStep)
5. Jigar Doshi (ARTPARK)
6. Pradeep Parappil (Megdap)
7. Richa Sharma (Wadhwani AI)
8. Sanyam Singh (Digital Green)
9. Syed Abdul Gaffar Shakhadri (Sandlogic)
10. Tahir Javed (AI4Bharat)
11. Trisha Singh (Wadhwani AI/Bhashini)
12. Vivek Seshadri (Karya)

Appendix 3

Objectives and Methodology

Appendix 3

Research Objectives

This research project examines the ecosystem of voice technologies in India, focusing on their development and deployment across technical, ethical, and legal dimensions. The objectives of our research are:

a. **Identify and analyse the key challenges** that arise across the lifecycle of voice technologies in India, including during data collection and curation, hosting, and downstream usage.

b. **Formulate actionable policy recommendations** grounded in both primary insights and secondary evidence, aimed at enabling the responsible development and use of voice technologies in Indian languages.

c. **Curate and present a set of best practices** for addressing the challenges identified, which can serve as a handbook for developers.

The study adopts an intersectional and lifecycle-based approach to comprehensively understand the challenges stakeholders, including data collectors and curators, model-hosting platforms, downstream developers, startups, and relevant government entities, encounter across technical, ethical, and legal dimensions.

Methodology

In addition to desk reviews, the team conducted qualitative interviews with stakeholders representing data collectors, curators, and model developers in India and globally.

a. **Multistakeholder dialogues:** A working group comprising representatives from key stakeholder groups, including government agencies, academic institutes, philanthropies, technology companies (both multinational and Indian startups), independent researchers, linguists, and civil society organisations working with open-source AI and speech data was established.

i. One-on-one interviews with working group members to gain an in-depth understanding of their perspectives on open source Indic voice technologies in India.

ii. Three workshop sprints conducted with expert groups to validate findings and gain additional insights on ethical, legal and technical issues.

b. **Secondary research**: A comprehensive review of existing scholarship and policy discourse on open source voice technologies in Indic languages, focusing on emerging opportunities, challenges, and mitigation strategies.

c. **Advisory board:** Establishment of a multidisciplinary advisory board to provide guidance on issues of open source voice technologies in Indic languages and to validate research directions and findings.

About the project

In 2025, Bhashini and GIZ Fair Forward jointly steered the consortium consisting of Artpark@IISc, Digital Futures Lab and Trilegal to explore the voice technology landscape in India, focusing on their development and deployment across technical, ethical, and legal dimensions.

BHASHINI
Bhashini is a Government of India initiative under the National Language Translation Mission (NLTM), focused on building an AI-powered national public digital platform for Indian languages with an aim to make language and technology accessible to everyone.

GIZ AND FAIR FORWARD - ARTIFICIAL INTELLIGENCE FOR ALL
The Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH is a federal enterprise with more than 50 years experience in international cooperation. On behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ), GIZ implements the project “FAIR Forward - Artificial Intelligence for All” which strives for a more open, inclusive and sustainable approach to AI globally.

ARTPARK@IISC
Artpark is a startup incubation and accelerator program designed to facilitate the evolution of a startup from innovation to incubation. It enables entrepreneurs and researchers to take ideas from the labs to the market, by bridging the gap between research innovations and their application in solving day-to-day challenges, specifically in the AI and Robotics ecosystem.

DIGITAL FUTURES LAB
Digital Futures Lab is an independent, interdisciplinary research studio that studies the complex interplay between technology and society in India and the Majority World. DFL works to realise pathways toward equitable, safe and sustainable digital futures through evidence-based research, systematic foresight and public engagement.

TRILEGAL
Trilegal is a full-service law firm in India with over 25 years of experience. The firm advises a diverse set of clients including Fortune 500 companies, global investment funds, major Indian conglomerates, domestic and international banks, technology and media companies, family offices and high net-worth individuals.

BHASHINI

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

FAIR Forward

TRILEGAL

digital futures lab

ARTPARK
AI & Robotics Technology Park @ IISc

nasscom ai
(Industry Advisor)