BHASHINI

Building an Open and Responsible Voice Technology Ecosystem

Policy Recommendations for Digital Inclusion in India

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

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

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.

This policy report documents how collaborative efforts in curating voice datasets and developing voice applications can create an inclusive, open and representative ecosystem. The policy recommendations address critical issues such as development of open-source foundational language datasets, purpose-specific customisation, consent considerations, copyright exemptions, value-sharing, and digital inclusion for India’s diverse Indic languages.

I encourage you to review the report 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

India’s remarkable linguistic diversity presents both opportunities and challenges for the development of inclusive voice technologies, shaping how millions can participate in the digital economy.

This report examines key barriers to building open and responsible speech systems in India—from data collection and model development to infrastructure and responsible practices. It proposes policy recommendations and governance mechanisms to support an innovative and equitable voice-technology ecosystem.

Key Recommendations

a. **Treating Foundational Datasets as Public Goods:** Foundational datasets for speech technologies are large, reusable corpora of audio, text, and metadata. They are curated to support a wide array of downstream applications, including automatic speech recognition (ASR), text-to-speech (TTS), and speech translation. Making foundational datasets available as digital public goods addresses market failures in voice-technology ecosystems and promotes local economic innovation.

   The report identifies challenges arising from the linguistic diversity and nuance of Indian languages, infrastructural barriers, the absence of common data standards that create governance gaps, and unresolved intellectual property and privacy issues. Policy recommendations for building foundational language datasets include clarifying and revisiting existing laws to enable the use of publicly available material, ensuring sustainable investments supported by government and blended finance techniques, and instituting strong governance systems with shared standards, coordinated repositories, and independent quality assurance.

   Making foundational datasets available as digital public goods addresses market failures in voice-technology ecosystems and promotes local economic innovation.

b. **Building Open and Representative Models:** India's success in inclusive voice technologies depends on whether speech systems perform well across the country's linguistic and social diversity. Today, limited openly available datasets, inadequate evaluation benchmarks, and uneven access to compute have resulted in models that perform inconsistently across languages, accents, and demographic groups, especially those from rural, low-resource, or marginalised communities. Strengthening India's ecosystem requires further investment in data, evaluation, and infrastructure. Benchmarking should be built around open evaluation datasets and transparent leaderboards providing common baselines for developers, improving procurement standards, and helping assess performance across diverse languages and speaker groups. On the infrastructure side, platforms such as Bhashini, ULCA, and AI Kosh offer strong foundations although effective operation hinges on sustained governance, clear access protocols, and long-term funding models.

Benchmarking should be built around open evaluation datasets and transparent leaderboards.

c. **Institutionalising Sustainable Open-Source Infrastructure:** Speech datasets place far greater demands on storage, bandwidth, and compute than text data, making the financing and governance of long-term hosting a central challenge for India's voice-tech ecosystem. A fragmented licensing landscape adds further uncertainty: overlapping or incompatible terms across data, code, model weights, and evaluation sets impose substantial compliance burdens on small actors, while enforcement gaps allow misuse with little recourse. Sustaining open, equitable development requires treating dataset hosting as durable public digital infrastructure rather than grant-based, project-specific assets. By international standards, India's emerging platforms, such as AI Kosh, provide a remarkable foundation. However, they require long-term funding, transparent governance, and clear access pathways for non-government contributors. Collaborative stewardship models, such as the Mozilla Data Collective, can help establish shared quality norms and consistent metadata conventions.

Sustaining open, equitable development requires treating dataset hosting as durable public digital infrastructure rather than grant-based, project-specific assets.

d. **Strengthening Responsible Deployment:** Deploying speech technologies responsibly requires more than high-performance models; it depends on safe systems, contextually appropriate use, and clear accountability. Existing data practices lack value-sharing mechanisms, leaving communities and researchers without recognition or benefit-sharing, even as their contributions fuel commercial products. Risks of misuse, including voice cloning, phishing, and deepfake-driven misinformation, are rising, and unintended harms like linguistic exclusion, biased performance across accents and genders, and the erosion of regional language identities remain widespread. Addressing these gaps requires structural guardrails that embed fairness, transparency, and accountability into deployment workflows.

Value-sharing mechanisms can help counter extractive data practices through attribution norms, community benefit-sharing, and the use of copyleft licences for publicly funded datasets.

Preventing misuse demands a combination of technical safeguards, stronger legal pathways, and widespread public literacy efforts to help users recognise risks and exercise their rights.

The report argues that a strong ecosystem requires more than innovation funding. Building open-source foundations—including language datasets, standards, collection protocols and responsible AI frameworks—promotes demand-driven local innovation. It is therefore essential that the state plays an active, shaping role, much as it has in the development of digital public infrastructure. In the context of voice technology, this involves both investing in commercially viable languages and sustaining low-resource languages that are vital for inclusion but unlikely to attract private capital. Open-source assets can reduce costs for the public and private sector alike. However, they demand long-term planning and financing for hosting, maintenance, and updates. These assets can be supported through blended-finance models that pool public, philanthropic, and commercial resources. Emerging national initiatives, such as the proposed AI marketplaces, can further structure participation, transparency, and value-sharing across data, annotation, and deployable models.

1. Role of Voice Technology in Building India’s Digital Future

1. Role of Voice Technology in Building India’s Digital Future

Voice technologies can be integrated into many existing digital services to open them to a wider range of users and make information more accessible. In doing so, they expand avenues for innovation and inclusion.[1] At their most basic level, they enable people to access information in their own language or even interact with digital applications using voice commands.

Imagine a farmer with access to up-to-date online information in their own language, including satellite data, meteorological patterns, and pest-control information. Or a student engaging with educational content in their mother tongue. Or an elderly person with failing eyesight communicating with family and navigating the digital world without reading or writing. Voice technologies help people overcome accessibility barriers, particularly those with limited digital literacy, low reading proficiency, or visual impairments.[2]

India is home to an extraordinary linguistic diversity. The 2011 census data documented 121 languages in the country, of which 22 are a part of the Eighth Schedule of the Constitution, and 99 are classified as other languages.[3] These figures do not capture many low-resource languages - languages with limited linguistic data, technical tools or resources for natural language processing.[4] While this diversity reflects our cultural strength, it also poses a significant technological challenge to inclusive digital access.

India’s digital ecosystem is expanding rapidly. By 2030, the country is projected to have about 1.2 billion smartphone users,[5] many of whom will speak local languages with limited online representation. Voice technologies can enable comprehensive and equitable digital access for these populations. The promise is not theoretical. India is already showing tangible progress, with the rapid proliferation of AI-enabled

[1] Interview with Venkatesh Hariharan, India Representative, Open Invention Network, virtual, 14 August 2025.

[2] Centre for Internet and Society (CIS), “Making Voices Heard: Policy Brief,” 2022, https://voice.cis-india.org.

[3] Government of India, “Census 2011, Language”, 2018, https://censusindia.gov.in.

[4] Low-resource languages are languages with limited linguistic data, technical tools and resources for natural language processing tasks like translation and text processing.

[5] GSMA Intelligence, The Mobile Economy Asia Pacific 2024 (London: GSMA, 2024), https://www.gsma.com.

voice applications such as speech recognition, generation, translation, and transcription services.[6] Bhashini, the Government of India's flagship initiative under the National Language Translation Mission (NLTM), has been established to provide the public access to voice technologies in multiple Indian languages.[7]

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 examine the challenges encountered by stakeholders—including data collectors and curators, model-hosting platforms, downstream developers, startups, and relevant government entities—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 was established 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.

[6] For example, Shivam Saxena, "Why NBFCs are Betting Big on Vernacular Voice AI," Gnani.ai, 2025, https://www.gnani.ai.
[7] "About Bhashini," Bhashini, accessed July 2025, https://bhashini.gov.in.

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

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

b. **Secondary Research:** A comprehensive review was undertaken 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:** A multidisciplinary advisory board was established to provide guidance on issues of open source voice technologies in Indic languages and to validate research directions and findings.

This report focuses specifically on open and responsible voice technologies, including speech datasets, models, and tools that are publicly accessible, transparently documented, and available for reuse by researchers, startups, civil society organisations, and government agencies. Openness is not merely a technical preference; it is a strategic approach for linguistic inclusion at India's scale. Proprietary, closed systems risk prioritising commercially viable languages while neglecting linguistic diversity, and may concentrate power in the hands of a few large corporations. Open approaches, by contrast, enable distributed innovation. When datasets and models are open, they become public digital goods, foundational resources that anyone can build upon, helping ensure that communities are not excluded simply because serving them is not profitable.

The report is structured as follows — **Section 2: Foundational Datasets as Public Goods** examines why datasets should be treated as public goods to address market failures, expand access, and promote local innovation. **Section 3: Building Open and Responsible Models** outlines structural barriers limiting open model development and use, and identifies steps to bridge these gaps. **Section 4: Institutionalising Sustainable Open-Source Infrastructure** examines current capacities, identifies persistent gaps, and highlights opportunities for long-term, reliable access. **Section 5: Responsible Deployment** details measures to ensure systems are safe, context-appropriate, and accountable in public and commercial settings. **Section 6:** Concludes the report.

2. Treating Foundational Datasets as Public Goods

2. Treating Foundational Datasets as Public Goods

Foundational datasets for speech technologies are large and reusable corpora of audio, text and metadata curated to support a wide array of downstream applications, including automatic speech recognition (ASR), text-to-speech (TTS), and speech translation. Unlike narrow datasets built for specific use cases, foundational datasets prioritise breadth to ensure representative coverage across language, dialect, geography, domain, speaker demographics, and recording conditions, enabling their use in both training and evaluation. Such datasets are useful for population-scale applications and designed to support generalisability. Developers can draw on these wide corpora and fine-tune them for domain- or context-specific applications.

> Foundational datasets prioritise breadth to ensure representative coverage across language, dialect, geography, domain, speaker demographics, and recording conditions, enabling their use in both training and evaluation.

Recognising foundational datasets as a public good helps address market failures in voice-technology ecosystems and promote local innovation. A useful parallel can be found in the case of Aadhaar, where base identity information enables interoperability across services like banking. In the context of speech data, large-scale, representative data is expensive to collect, and private actors are not incentivised to invest in foundational datasets when use-case-specific datasets serve their needs. While widely-spoken languages such as Hindi and its dialects might attract private investment, languages spoken by smaller populations face structural neglect. Private actors rationally focus on high-return languages, creating a long tail of linguistic exclusion that public investment can address. This is not merely an equity concern but also an efficiency concern: without comprehensive linguistic coverage, voice technologies cannot fulfil their potential as digital inclusion tools.

India has been promoting initiatives that develop institutional architecture for open speech data.[8] However, many challenges persist. While some stem from infrastructural constraints, others arise from the peculiar nature of Indian languages. The following section examines the current landscape of India's open speech dataset ecosystem and the systemic challenges that shape it.

[8] For instance, Unified Language Contribution API is an open scalable data platform within Bhashini for standardising all data and model contributions for benchmarking, please see Bhashini. Similarly, AI Kosh provides a repository of Open Sourcedatasets and models along with compute infrastructure, tiered access protocols, and a sandboxed environment for developers and researchers. See also "AI Kosh," IndiaAI, https://aikosh.indiaai.gov.in/.

2.1 Challenges in Building Foundational Dataset Ecosystem

Government-supported initiatives in India, such as IndicVoices[9], Project Vaani[10], Speech Recognition in Agriculture and Finance for the Poor in India (ReSPIN)[11], and SYSPIN[12], are developing datasets for key Indian languages with explicit attention to representativeness and dataset quality. These initiatives lay a strong foundation for India's open voice ecosystem and demonstrate the feasibility of large-scale, responsible data collection practices.

The challenges facing India's speech dataset ecosystem can be organised into the following interconnected categories. First, technical and quality challenges arise from the country's exceptional linguistic diversity and the uneven distribution of infrastructure. Second, governance gaps reflect the absence of unified standards for data management, annotation, and evaluation. Third, legal concerns persist due to ambiguities in copyright and data protection laws. These challenges, detailed below, limit the scalability, reusability, and trustworthiness of Indian speech datasets.

a. **Speech Data Collection:** India's linguistic landscape is exceptionally diverse, comprising over 19,500 languages and dialects,[13] each with unique linguistic features and cultural contexts. Documenting this diversity at scale is both technically and socially complex, particularly when developing inclusive datasets. Linguistic variation and nuance further complicate data collection processes. The challenges affecting representativeness and data quality can be broadly categorised into three areas: linguistic nuances of Indian languages, workforce challenges, and issues with data collection techniques:

   - **Linguistic nuances:** Current English-centric preprocessing techniques (e.g., tokenisation) and language models disregard the linguistic diversity, morphological complexity, and dynamic evolution

[9] Tahir Javed et al., "Lahaja: A Robust Multi-Accent Benchmark for Evaluating Hindi ASR Systems," preprint, arXiv:2408.11440 (2024), https://arxiv.org.
[10] Artpark, "Capturing the Language Landscape for an Inclusive Digital India", 2025, https://vaani.iisc.ac.in.
[11] RESPIN, Indian Institute of Science, https://respin.iisc.ac.in.
[12] SYSPIN, Indian Institute of Science, https://syspin.iisc.ac.in.
[13] Government of India, "Census 2011, Language", 2018, https://censusindia.gov.in/

of Indian languages, which are often absent in English.[14] In addition, primarily oral languages that do not have standardised written scripts are often invisibilised because they do not fit into neat taxonomies of orthography. For example, a 2024 work on the Gormati language found that verbatim translation of written language ideas does not work for oral languages.[15]

- **Shortage of qualified transcribers and annotators:** A scarcity of skilled transcribers, especially those fluent in low-resource languages and trained in standardised annotation processes, makes it difficult to ensure adequate representation, even with government support. Researchers associated with Project Vaani note that economic issues compound these technical and human resource challenges, as the cognitively demanding, intricate nature of transcribing low-resource Indian languages makes it a less lucrative profession, discouraging potential workers. These challenges are further exacerbated by the fragmented nature of the labelling industry, where the absence of formal channels to connect data users with individual labellers often results in a significant disparity between what data users pay and the compensation workers receive.

- **Issues with data collection techniques:** Datasets may be distorted or unnatural in many ways. Issues like multi-speaker overlap, background noise, volume inconsistencies, or poor articulation reduce intelligibility and limit the dataset's usefulness in training reliable models.[16] A single person dominating the conversation, off-topic responses in domain-specific interactions, mispronunciations, and echoes caused by recording environments can further impact the audio quality.[17] Additional challenges arise from content that includes profanity, sensitive topics, or frequent code-switching between languages, which may not always be annotated or handled appropriately.[18]

b. **Lack of Standards and Guidelines for Data Management:** The dataset ecosystem in India is fragmented across research groups and public initiatives. The absence of frameworks limits reusability and results in uneven dataset quality across projects, institutions, and states. These gaps make it difficult to create AI-ready corpora that can be combined, benchmarked, or scaled across use cases.

[14] Farhana Shahid, Mona Elswah, and Aditya Vashistha, "Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages," preprint, arXiv:2501.13836 (2025), https://arxiv.org/html.

[15] Thomas Reitmaier 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] Ali Sartaz Khan et al., "Multi-Stage Speaker Diarization for Noisy Classrooms," preprint, arXiv:2505.10879 (2025)

[17] Interview with Janki Nawale, Linguist/Researcher, AI4Bharat', virtual, 20 June 2025.

[18] Bharathi Raja Chakravarthi et al., "Dravidiancodemix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text," Language Resources and Evaluation 56, no. 3 (2022): 765–806, https://link.springer.com

Several challenges contribute to this problem, including:

- **Inconsistent metadata practices:** Different data collection entities have different standards for metadata collection. For example, while some entities collect only minimal speaker attributes, others collect information on dialect, socio-economic background, literacy, audio environment, and prompt type. This lack of consistency across datasets makes it challenging to evaluate dataset diversity, track biases, or assess model performance in real-world scenarios.

- **Divergent annotation conventions:** Poor annotation and labelling critically affect model performance and data utility. Inaccurate or inconsistent labelling, transcription errors, incorrect speaker tags, or misclassified language can all introduce noise into the data. This can impact the model performance, compromise the effectiveness of downstream applications, and hinder reliable model comparison.[19] This issue is compounded by the lack of qualified transcribers who are both native language speakers and proficient writers familiar with the spelling and grammatical structures of these languages.[20] The scarcity of skilled labellers makes it difficult to ensure high-quality labelling at scale. Consequently, subjectivity in transcription is common, leading to inconsistent data.

- **Absence of evaluation frameworks suited for Indian languages:** The lack of standardised methods for assessing datasets in Indian languages hinders the development of robust voice technologies, exacerbating the performance gap relative to Western languages. Existing dataset evaluation frameworks are heavily skewed toward Western linguistic contexts. Researchers argue that metrics like Word Error Rate (WER) and Character Error Rate (CER) are inadequate for evaluating dataset quality in most of the Indian languages, as they fail to account for linguistic complexities like code-switching, tonal variations, and cultural nuance.[21,22,23]

c. **Legal Ambiguities in Data Collection and Curation Practices:** The legal environment governing speech datasets in India is complex and fragmented, creating substantial uncertainty for researchers, companies, and public institutions working with speech data. Issues related to copyright, data protection, content regulations, and any restrictions under platform-level documentation must be accounted for.

[19] Label Studio, "6 Costly Data Labeling Mistakes and How To Avoid Them," September 2022, https://labelstud.io/

[20] Interview with Aaditeshwar Seth, Aaditeshwar Seth, Professor, Department of Computer Science and Engineering, IIT Delhi, virtual, 24 June 2025.

[21] Anushka Singh et al., "How Good Is Zero-Shot MT Evaluation for Low Resource Indian Languages?," preprint, arXiv:2406.03893 (2024)

[22] Laurent Besacier et. al. Automatic speech recognition for under-resourced languages: A survey

[23] Ashwani Tanwar et. al. Translating Morphologically Rich Indian Languages under Zero-Resource Conditions

The challenges fall into three broad categories:

- **Overlapping copyright layers and unclear licensing terms:** Under Indian law, particularly the Copyright Act, 1957, multiple layers of intellectual property protection may apply: to underlying text or transcripts (as literary works[24]), voice recordings (as sound recordings[25]), and curated metadata, provided each meets originality requirements. Using 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, because such material lacks originality, and is therefore free to use without restriction.

  Furthermore, crowdsourced datasets may lack clarity about contributors' rights, particularly when the terms of data collection are not clearly communicated. With thousands of speakers involved, it is often practically challenging to obtain clarification or additional permissions retroactively, which could, in some cases, also block open release. The Baidu Deep Speech case[26] illustrates how even well-intentioned crowdsourcing efforts can create insurmountable licensing gaps where creators may be unable to identify the contributors, resulting in an orphan works situation.[27] A legal review revealed that Baidu's original agreements with 9,600 speakers covered internal use but explicitly excluded public release or unrestricted commercial reuse. However, it was practically impossible to contact thousands of contributors to secure broader licensing terms retroactively. This case underscores how complications can arise not only from third-party content, but also from inadequate initial rights clearance during the data collection process. This leads to orphan works situations in which the intended use exceeds the scope of granted permissions and remediation becomes prohibitively complex.

  On a related note, the Indian judiciary is actively addressing issues involving AI systems, with emerging jurisprudence, particularly in IP injunction suits. Recent rulings on the non-consensual use of an individual's likeness, name, image, and voice suggest evolving principles that treat AI-generated deepfakes exploiting celebrity personas as violations of privacy and personality rights.[28]

[24] Section 2(o), (Indian) Copyright Act, 1957

[25] Section 2(xx), (Indian) Copyright Act, 1957

[26] Galvez et al., The People's Speech.

[27] U.S. Copyright Office, Orphan Works and Mass Digitization: A Report of the Register of Copyrights, (Washington DC: U.S. Copyright Office, June 2015), https://www.copyright.gov/

[28] Anil Kapoor v. Simply Life India and Ors., CS (COMM) 652/2023 and I.A. 18237/2023-18243/2023; Arijit Singh v. Codible Ventures LLP, Interim Application (L) No.23560 of 2024 in Com IPR Suit (L) No.23443 of 2024.

- **Risks arising from secondary or scraped data regarding unlawful content:** When speech datasets are sourced from secondary materials, particularly through web scraping or large-scale crawling, there is a risk that data may be collected from contexts involving sensitive or high-risk content. Audio may be extracted from online video-hosting platforms, podcasts or even radio broadcasts, capturing a range of languages, dialects, accents and contexts.

  For example, scraped data may include content from platforms hosting sexually explicit material, forums discussing personal or health-related experiences, or sources that feature the voices of minors. If such content is not properly segregated or flagged during curation, it may be incorporated into foundational speech datasets intended for broad downstream use, raising significant legal, ethical, and reputational risks. This risk is compounded when scraped data lacks reliable metadata or provenance, making it challenging to identify and exclude harmful content.In such cases, the risk of attracting legal liability under existing content regulations cannot be ruled out, particularly given that current legal principles may not clearly apportion liability on the relevant actor in the AI value chain.

- **Consent and privacy complexities under data protection laws:** India's DPDP Act adopts a consent-centric model. All processing of personal data—regardless of its sensitivity—requires valid consent, with limited grounds for non-consensual processing. The Act recognises certain legitimate uses where consent is not required, such as for state functions, medical emergencies, or to fulfil legal obligations. However, these are narrowly defined and exclude most commercial applications. Under the DPDP Act, 2023, data protection obligations do not apply to personal data that the data principal has made publicly available themselves, or that any other person is legally required to disclose publicly.The DPDP Act also provides exemptions for research, archival, or statistical purposes, provided such personal data is not used to make a decision about a specific data principal, and that prescribed technical and organisational safeguards are implemented under the Digital Personal Data Protection Rules, 2025 (DPDP Rules).

  Decisions made during data collection and curation regarding intellectual property and data protection safeguards are critical, as they determine how a dataset can be used and shared. However, verifying that a data principal was in fact the source of a disclosure can be a challenging task. This creates ambiguity between content that is merely"'publicly available" and that which has been made "public" by the individual, an issue that is particularly relevant when sourcing voice recordings or transcripts from online platforms and archives.

In a similar vein, there are practical challenges around relying on consent as the basis for processing personal data while also adhering to data minimisation and other such principles, especially given how AI models and systems operate. However, the applicability of the research exemptions in the context of AI training or commercial research activities (instead of relying on consent as the legal basis for processing) remain uncertain. How these interpretations will evolve in practice is yet to be determined.

As a separate aside, copyright related considerations continue to apply in relation to the use of publicly available data scraped from the web.

These gaps highlight the critical policy levers that must be activated to ensure equitable, high-quality, and sustainable data development. The following section proposes interventions across institutions, standards, and infrastructure.

2.2 Policy Recommendations

Addressing the challenges outlined above requires coordinated intervention across funding, governance, and infrastructure. The following recommendations are mutually reinforcing: revisiting exemptions under the intellectual property and data protection regimes can provide developers with clarity about acceptable use; public investment creates datasets; and governance frameworks ensure their quality and usability. Together, these measures recommendations would help establish speech data as durable public digital goods over the long term.

Our recommendations are mutually reinforcing: revisiting exemptions under the intellectual property and data protection regimes can provide developers with clarity about acceptable use; public investment creates datasets; and governance frameworks ensure their quality and usability.

a. **Exploring Exemptions Under the Copyright Act, 1957 and the DPDP Act, 2023** From a copyright standpoint, the multi-stakeholder committee constituted by the Department for Promotion of Industry and Internal Trade (DPIIT) to study the intersection of AI and copyright law has recently issued the Working Paper on Generative AI and Copyright Part 1 for stakeholder feedback. In assessing the use of copyright-protected works for AI model training, the committee evaluated various models and, based on a majority view, proposed a hybrid model. Notably, this hybrid model proposes a mandatory blanket license under the Copyright Act, 1957, enabling AI developers to use all lawfully acquired copyright-protected works for AI training without requiring prior permission from copyright owners. At the same time, the model establishes a statutory remuneration right for copyright holders, ensuring fair compensation through the establishment of the Copyright Royalties Collective for AI Training, a centralised, non-profit entity composed of rights-holder associations. Royalty rates are to be determined by a government-appointed Rate Setting Committee. The model also recommends retroactive application, obliging AI developers already commercialising systems trained on copyrighted content to pay royalties for past use.

While the proposed hybrid model for generative AI and copyright seeks to balance innovation with fair remuneration for rightsholders, it faces significant practical challenges.These include the operational complexity of setting fair royalty rates, accurately collecting revenue from global AI companies, and equitably distributing payments to rightsholders—particularly when a large volume of copyrighted work is unlikely to be registered—and other complications arising from retroactive application. However, it remains to be seen how these practical challenges would be addressed should the hybrid model proposed by the DPIIT be implemented.

From a data protection standpoint, certain key clarifications would support the development of AI technologies and help practitioners structure their compliance programs accordingly. Notably, guidance on how to determine whether any personal data has in fact been made publicly available, as well as on interpreting the contours of the research exemption, would be particularly beneficial.

Further, given the unique characteristics of AI technologies, it would be useful to assess the appropriateness of existing data protection principles (e.g., data minimisation) in light of how AI models and systems operate, and to explore creating a new, specific ground for processing tailored to AI technologies.

Additionally, where consent is relied on as the basis for processing personal data, ensuring that such consent is genuinely informed is critical, particularly when community participation is involved. In this regard, intermediaries can play a vital role. This includes entities and individuals who facilitate "last-mile" communication on the ground by translating notices and obtaining consent in local languages, as well as technological intermediaries—such as consent managers—that provide the digital infrastructure to manage and record consents in one place

Given India's vast linguistic diversity, these intermediaries are essential for bridging communication gaps by presenting consent requests in local languages and culturally relevant contexts. This ensures that consent is not merely a formality but is truly informed, thereby fostering trust and enabling meaningful participation. Incentivising such entities should therefore feature in India's policy roadmap.

b. **Fund Representative Foundational Dataset and Model Creation:** The government should adopt an explicit public-good orientation for foundational linguistic datasets, recognising their role as shared digital infrastructure that underpins equitable innovation. Bhashini is already funding multiple programs through academic institutions, with a focus on creating foundational datasets under open-source licenses to build language corpora in underserved regions. For instance, India is supporting domestic companies such as Sarvam AI, Soket AI, Gan AI, and Gnani AI in building indigenous foundation dataset models for Indian languages.[29]

Such interventions require dedicated public investment in long-term, large-scale, and representative data collection, led by national institutions in partnership with regional academic and community organisations. Funding mechanisms should prioritise the inclusion of underserved, low-resource, tribal, and oral languages, and ensure that resulting datasets are validated against rigorous quality standards. Publicly funded datasets should, by default, be open access and governed to maximise safe and fair reuse across research, social innovation, and local industry, subject to appropriate privacy, intellectual property, and access safeguards. To enable such an ecosystem, the National Roadmap for Artificial Intelligence by NITI Aayog proposes creating a national AI marketplace comprising a data marketplace, a data annotation marketplace, and a deployable model/solutions marketplace to encourage participation of individuals and communities in the voice technologies ecosystem.[30] The data marketplace is envisioned to be built on blockchain technology and to feature traceability, access controls, compliance with regulations, and robust price discovery mechanisms.[31]

Targeted mechanisms for continuous stakeholder engagement require a multi-pronged approach. Engaging educational institutions, such as engineering colleges, in a structured data collection effort could accelerate corpus development while building the necessary consent and quality control mechanisms. Contribution acknowledgements and fellowships can motivate individuals and community organisations to participate in data creation and annotation.[32] The value of such incentives was evident in the IndicVoices project. For many contributors under this initiative, the motivation extended beyond monetary compensation to include language preservation, and the knowledge that their work could improve communication technologies and daily digital experiences, such as understanding content on platforms like YouTube.[33]

[29] S Ronendra Singh, “Government Selects 3 Start-Ups to Develop Local Ai Models under India Ai Mission,” The Hindu Businessline, May 2025, https://www.thehindubusinessline.com

[30] NITI Aayog, National Strategy for Artificial Intelligence, (New Delhi, NITI Aayog,2018), https://www.niti.gov.in

[31] NITI Aayog, 2018.

[32] Francesco Cappa et al., “Bring them aboard: Rewarding Participation in Technology-Mediated Citizen Science Projects,” Computers in Human Behavior 89 (2018): 246-57, https://www.sciencedirect.com.

[33] Interview with Janki Nawale, Linguist/Researcher, AI4Bharat, virtual, 20 June 2025.

c. **Create Strong Governance and Quality Systems for Foundational Speech Data:** Institutional mechanisms are needed to ensure that high-quality dataset development is not fragmented or ad hoc. This includes establishing clear national standards for dataset documentation, governance, quality validation, version control, and responsible release. Dedicated coordination bodies that can be potentially housed within public institutions should steward shared repositories for datasets, language models, and tools, and should administer tiered access mechanisms. Independent review processes for evaluating dataset quality, representativeness, and risk safeguards would strengthen accountability and trust across the ecosystem, especially in deployments through public institutions.

Foundational speech datasets should be treated as public goods, supported by institutions that can maintain them over time. This would help ensure that voice technologies serve all language communities, and not only those profitable to private companies. Together, these recommendations offer a practical pathway from today's fragmented data efforts toward a coordinated, reliable, and long-term public infrastructure for speech technology.

3. Building Open and Representative Models

3. Building Open and Representative Models

India's ambition to build inclusive voice technologies hinges on whether speech systems can perform reliably across its linguistic, social, and regional diversity. Current ecosystem gaps rooted in uneven datasets, weak benchmarks, capacity constraints, and limited compute directly affect the performance, fairness, and real-world usability of open models built for inclusion. Without deliberate efforts to address these limitations, speech technologies risk reproducing existing patterns of societal exclusion. This section outlines these challenges, examines their impact on inclusive voice technologies, and offers suggestions to address them.

Current ecosystem gaps rooted in uneven datasets, weak benchmarks, capacity constraints, and limited compute directly affect the performance, fairness, and real-world usability of open models built for inclusion.

3.1 Challenges in building Open and Representative Models

a. **Lack of Shared Benchmarks for Evaluation:** Very few standardised, widely adopted benchmarks exist for Indian languages, significantly hindering the development and evaluation of language technologies. In contrast to high-resource languages like English, which benefit from mature and universally recognised benchmarks (e.g., GLUE for NLP or LibriSpeech for ASR), Indian languages lack comparable, domain-relevant, and well-curated evaluation datasets. Creating consistent benchmarks is thus essential for progress.[34] In their absence, it becomes difficult to objectively compare model performance across different systems, teams, or research initiatives. This gap has profound implications for both academia and industry. In research contexts, it affects reproducibility, as the lack of standard datasets and evaluation protocols makes it difficult to verify or build on previous work. In industry settings, it creates uncertainty in deployment decisions, as there are no reliable metrics or standard baselines to assess whether a model is production-ready or suitable for real-world use across multiple Indian languages and dialects.

b. **Uneven Compute Access:** Training state-of-the-art speech models demands substantial computational resources, including high-end GPUs, fast storage systems, and large memory capacity, all of which are costly to acquire and maintain. This poses a significant barrier, particularly in resource-constrained settings like academic institutions, startups, and developing regions. For example, training the Whisper model required approximately 680,000 GPU hours on 16 A100 GPUs over two months, highlighting the scale of computation involved.[35]

   In response to this challenge, the Government of India, through the Ministry of Electronics & IT, has announced the provision of over 18,000 affordable AI compute units under the IndiaAI Mission.[36]

[34] Hussam Azzuni and Abdulmotaleb El Saddik, “Voice Cloning: Comprehensive Survey,” preprint, arXiv:2505.00579 (2025), https://arxiv.org.

[35] Alec Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” in Proceedings of the International Conference on Machine Learning (PMLR, 2023), 28,492–28,518.

[36] IndiaAI, “IndiaAi Compute Capacity,” 2025, accessed 31 July, 2025, https://indiaai.gov.in.

Eligible users can access these resources at discounts of up to 40 percent. This initiative marks a significant step toward democratising access to AI infrastructure, fostering broader participation in AI research and development, and strengthening India's position in the global AI landscape.

Taken together, these constraints shape who can meaningfully build, evaluate, and deploy speech models in India, and who remains excluded from their benefits. Addressing them requires coordinated action across data, evaluation, compute, and transparency. The following recommendations outline practical steps that institutions, researchers, and government actors can take to strengthen the ecosystem and support the development of open and representative voice technologies.

3.2 Policy Recommendations

a. **Create Publicly Accessible Evaluation Datasets and Leaderboards for Indian Languages:** Indian developers currently lack standardised ways to assess whether voice technologies actually work for Indian speakers. This creates challenges for public procurement and performance measurement. For instance, in the absence of a standardised benchmarking, government departments have no objective criteria to evaluate vendor claims. Furthermore, developers may struggle to identify shortcomings when systems tested in labs fail in real-world deployment.

To address the fragmentation in evaluation practices, a nationally coordinated benchmarking initiative is needed. Public evaluation datasets should be openly released under transparent licensing terms, and designed to cover real-world conditions such as conversational speech, code-switching, noisy environments, dialectal variation, domain-specific terminology, and different demographic groups. These evaluation datasets should be regularly updated to avoid gaming of the system. A corresponding leaderboard can provide comparable performance reporting, incentivising high-quality submissions, and enable auditing of accuracy across languages and speaker groups.

Public evaluation datasets should be openly released under transparent licensing terms, and designed to cover real-world conditions such as conversational speech, code-switching, noisy environments, dialectal variation, domain-specific terminology, and different demographic groups.

Similar initiatives in India and abroad show that coordinated, open benchmarking is both feasible and valuable. In the Indian context, efforts like IndicSUPERB[37] and Vistaar[38] already demonstrate the usefulness of publicly released evaluation datasets and open scoring pipelines, even though their coverage of dialectal and real-world

[37] Tahir Javed et al., "Indicsuperb: A Speech Processing Universal Performance Benchmark for Indian Languages," In Proceedings of the AAAI Conference on Artificial Intelligence, 37, no. 11 (2023), 12942-50, https://ojs.aaai.org.

[38] AI4Bharat, "Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR," GitHub repository, https://github.com

variability remains limited. Internationally, multilingual suites such as ML-SUPERB[39] and community-maintained platforms like the Open ASR Leaderboard[40] illustrate how shared benchmarks and transparent leaderboards can enable stronger comparability across systems.

To operationalise evaluation benchmarks in the Indian context, a coordinated agency of national and sub-national bodies can be established. Such an entity can develop and maintain evaluation benchmarks across India's linguistic diversity. Sub-national bodies should be enabled to create evaluation datasets for their respective languages and priority use cases, while a Union-level entity such as Bhashini should serve as the central convener and standard-setting body, including supporting the development of benchmarks for low-resourced languages.

Taken together, these examples highlight the value of structured, publicly governed benchmarking ecosystems: reproducible evaluation, independent verification of vendor claims, and clearer performance standards for developers. At the same time, they reveal the gaps that a government entity could fix. These gaps include skewness toward high-resource languages, clean audio, or narrow demographic profiles. A national initiative in India would need to address these shortcomings directly by accounting for conversational code-switching, regional dialects, varied noise conditions, and demographic diversity across states. By building on both domestic and international precedents while addressing their limitations, India can establish a public evaluation infrastructure that reflects real usage conditions and strengthens procurement, accountability, and ecosystem-wide quality.

b. **Pool Public and Academic Compute Resources to Lower Barriers for Training and Fine-Tuning:** A sustainable strategy for foundational datasets must be paired with accessible and affordable compute and storage, particularly for publicly funded institutions, research centres, and community-led language projects. This includes establishing shared national compute and storage facilities with subsidised access, transparent allocation rules, and support for training and onboarding. Preferential access can be provided to projects committed to open-sourcing their code, model and/or datasets. Strengthening regional university capacity with on-site infrastructure prevents centralisation and enables data processing, annotation, and model experimentation closer to the communities where data is collected. In parallel, dedicated funding and procurement pathways are needed to allow for secure

[39] Jiatong Shi et al., "Ml-superb 2.0: Benchmarking Multilingual Speech Models across Modeling Constraints, Languages, and Datasets," arXiv preprint arXiv:2406.08641 (2024), https://arxiv.org.

[40] "Open ASR Leaderboard," Hugging Face Spaces, https://huggingface.co.

storage compliant with data protection requirements. Democratising access to compute is essential to enable innovation beyond a small circle of corporate or elite research institutions.

India has made notable strides in building infrastructure to support voice technology through platforms such as Bhashini and AI Kosh. Within Bhashini, the Unified Language Contribution API (ULCA) serves as a standard API and open scalable data platform designed to standardise all data and model contributions for benchmarking.[41] AI Kosh, launched as part of the IndiaAI Mission in March 2025, provides a repository of open datasets and models alongside compute infrastructure, tiered access protocols, and a sandboxed environment for developers and researchers.[42]

India's planned provision of subsidised compute under the IndiaAI Mission is a strong start. However, it should be complemented by shared academic clusters, government-supported compute credits, and structured residency programmes that enable startups, student researchers, and community organisations to undertake compute-intensive training or fine-tuning.

[41] "Unified Language Contribution API," Bhashini, https://bhashini.gov.in.
[42] "AIKosh," IndiaAI, https://aikosh.indiaai.gov.in/home.

4. Institutionalising Sustainable Open Source Infrastructure

4. Institutionalising Sustainable Open Source Infrastructure

Speech datasets require considerably more storage per sentence and computational resources than text datasets, creating unique sustainability challenges for open source infrastructure. For instance, AudioSet totals nearly 2.5TB and contains approximately 52 million spoken words. By contrast, the C4 dataset[43], a text-only corpus, is about 30 percent of AudioSet's file size while containing 153 billion words.[44] This disparity reflects the inherent information density of speech datasets, which include indicators such as accent, emotion, speech rate, and other such features that accompany the speech recording.

The challenges of building open and representative speech datasets thus extend beyond data collection and model training; they also arise from how data is hosted, shared, and governed. Maintaining large speech datasets requires sustained financial investment, yet long-term hosting remains precarious for many academic groups, civil society organisations, and smaller developers who depend on free or subsidised storage options with uncertain continuity. Without stable, accessible, and well-governed infrastructure, even high-quality datasets risk becoming inaccessible resources.

[43] Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research 21, no. 140 (2020): 1–67.

[44] William Agnew et al., "Sound Check: Auditing Audio Datasets," preprint, arXiv:2410.13114 (2024), https://arxiv.org.

4.1 Challenges of Institutionalising Sustainable Open Source Infrastructure

In India's voice-technology ecosystem, hosting entities — such as public repositories or research platforms—bear the costs of long-term cloud storage and bandwidth, as well as the responsibility for instituting a governance framework for these datasets. While the Government of India is seeking to address some of these barriers through initiatives such as AI Kosh, which aims to host AI datasets and models,[45] the platform is currently nascent, and its utility for the voice technology sector is yet to be demonstrated. Early indications suggest promising infrastructure investments, but questions remain regarding long-term funding commitments and access policies for non-government actors.

Financial fragility, weak interoperability, and legal ambiguity hinder the reliable reuse, maintenance, and equitable development of voice technologies.

Moreover, the absence of consistent interoperability practices compounds these structural constraints. Most hosting platforms do not enforce standard tagging, provenance tracing, or unified upload standards, making datasets difficult to discover, verify, or reuse, particularly for teams operating with limited technical capacity. This complexity is exacerbated by a fragmented licensing landscape: speech datasets, model weights, code, and evaluation sets often come with overlapping or incompatible licenses, and enforcing permitted-use restrictions across jurisdictions or downstream users is rarely feasible. These issues create an ecosystem in which financial fragility, weak interoperability, and legal ambiguity hinder the reliable reuse, maintenance, and equitable development of voice technologies.

a. **Financial Sustainability of Hosting:** The size and complexity of speech datasets mean they require substantial cloud storage and bandwidth,

[45] Ministry of Electronics and Information Technology, "MeitY Launches AIKosha, a Secured Platform That Provides a Repository of Datasets, Models and Use Cases to Enable AI Innovation," press release, 2025, https://www.pib.gov.in.

costs that most academic institutions, civil society groups, and smaller developers cannot easily afford. Many rely on third-party services to host their data, leading to long-term sustainability concerns, especially when funding cycles end or priorities shift. While some hosting entities allow free public hosting of open source datasets, the uncertainty around the durability of such arrangement remains a concern for data collectors and developers.[46] Academic institutions, which conduct much of India's foundational research in voice technology, typically lack dedicated line items for long-term data hosting. Project grants fund initial collection and curation but rarely include provisions for hosting beyond the grant period. Once funding ends, datasets face an uncertain future: costs may be borne personally by the principal investigator (unsustainable), the institution may absorb costs (rarely formalised), or the dataset is migrated to free platforms with uncertain longevity.

Free or subsidised hosting platforms such as Hugging Face[47] and Zenodo[48] partially mitigate these costs but introduce different vulnerabilities. These platforms depend on philanthropic funding, venture capital, or university budgets, none of which guarantee long-term availability.

b. **Issues with Interoperability and Reuse:** Hosting entities do not typically enforce standardised tagging and precise data provenance requirements.[49] This makes it difficult for developers, especially those working in low-resource environments, to identify and reuse data effectively. Inconsistent hosting practices, such as duplicative or incomplete uploads across platforms like GitHub and Hugging Face, further complicate dataset discoverability.[50] These barriers disproportionately affect smaller teams, academic researchers, and grassroots organisations, who often lack the technical capacity to navigate a disorganised data landscape, consequently reinforcing existing inequities in voice technology development.

The absence of standardised metadata provenance creates obstacles for downstream developers and researchers, who must download and inspect datasets to determine whether they are useful. This process is time-consuming and can become prohibitive for small teams. Provenance tracking fails when datasets are forked, modified and reuploaded without clear documentation of changes, making it

[46] Alex Gude, "Where to Host Public Datasets?," https://alexgude.com.
[47] Hugging Face, https://huggingface.co.
[48] Zenodo, https://zenodo.org.
[49] Edd Gent, "Public AI Training Datasets Are Rife With Licensing Errors An audit of popular datasets suggests developers face legal and ethical risks", IEEE Spectrum, 8 November 2023, https://spectrum.ieee.org.
[50] Hugging Face, https://huggingface.co and GitHub, https://github.com.

challenging to verify data integrity or trace errors to their source. Such issues can create significant entry barriers for small players.

c. **Complexities in Licensing:** Licensing frameworks clarify usage rights by explicitly stating what licensors may or may not do with AI system components. However, their effectiveness is often constrained by a range of practical and systemic challenges. For example, enforcement remains a significant hurdle; even when licenses explicitly restrict specific uses, enforcing these terms, especially across jurisdictions or with downstream users, is complex and often impractical.51 In addition, AI systems comprise multiple components—source code, training data, model weights, and evaluation data—each potentially governed by a different license. These overlapping and sometimes conflicting terms create significant challenges for harmonisation, particularly when pre-trained models are reused or fine-tuned. The cumulative effect is legal uncertainty and increased compliance costs, which disproportionately affects smaller actors.[52]

   Enforcement presents even greater challenges. For example, a research group may release a speech dataset under a CC-BY-SA licence that requires attribution and share-alike. If a downstream developer uses it in a proprietary commercial voice assistant without honouring these terms, the research group currently has no meaningful recourse. This enforcement gap creates perverse incentives. Well-intentioned actors invest time in compliance, carefully tracking licence obligations and releasing derivative works appropriately.

   On the other hand, bad actors may simply ignore license terms. The result is a “tragedy of the commons” dynamic, in which open datasets and models are exploited without due reciprocal contribution. This can undermine the sustainability of open source ecosystems and discourage future contributions.

d. **Version Control and Data Provenance:** Version control and dataset evolution present additional challenges for speech datasets. Unlike software code, where version control systems like Git provide robust change tracking, speech datasets lack standardised versioning practices. Datasets evolve as errors are corrected, more speakers are added, annotation improves, or quality enhancements are made. However, these changes are poorly documented and tracked. As a result, downstream users are unable to access the most current,

[51] Kevin Klyman, “Acceptable Use Policies for Foundation Model’,” preprint, arXiv:2409.09041, 29 August 2024, https://doi.org.
[52] Moming Duan et al., “ModelGo: A Practical Tool for Machine Learning License Analysis,” in Proceedings of the ACM Web Conference 2024, (New York: ACM, ACM, 13 May 2024), 1158–69, https://doi.org.

accurate open datasets and models.[53] Model benchmarks become incomparable when different teams unknowingly use different dataset versions. Errors in datasets, such as misclassified speakers or incorrect transcriptions, may be identified and corrected, but without explicit versioning and changelogs, these corrections propagate slowly and inconsistently across the ecosystem.

This problem is further exacerbated by “dataset drift”: the gradual, undocumented evolution of datasets over time as hosting platforms compress files, convert formats, or apply processing pipelines. Metadata might be stripped during format conversion. These changes accumulate silently, creating situations in which what is ostensibly the “same” dataset differs substantively on different platforms or at varying points in time.

[53] Interview with Soma Dhavala, IIT Jammu, virtual, 24 September 2025.

4.2 Policy Recommendations

India requires dedicated, sustainably funded infrastructure for hosting speech datasets as an enduring public digital good, rather than as project-dependent resources. AI Kosh represents a promising start; however, a comprehensive approach requires institutional commitment, transparent governance, and guaranteed long-term funding. In addition, to improve discoverability and interoperability in open speech datasets and models, standardised open documentation practices should be embedded across entities through a collaborative data stewardship model.

A sustainable open source infrastructure requires institutional commitment, transparent governance, and guaranteed long-term funding.

a. **Embed Standardised Open Documentation Practices Across Projects:** Consistent documentation through data cards, model cards, version logs, and provenance statements is critical for transparency, interoperability, and long-term maintainability. Publicly funded datasets and models should follow compulsory documentation norms that specify collection methodology, demographic composition, known limitations, ethical safeguards, and usage constraints. Mandating these standards at the project approval or dissemination stage would enable future researchers, developers, and community organisations to evaluate dataset fitness, reproduce results, and identify gaps more effectively.

b. **Adopt Collaborative Data Stewardship Models:** The Mozilla Data Collective[54] demonstrates how distributed contributors can pool datasets under clear licensing terms, shared metadata conventions, and community-driven governance. Under such models, data contributors—whether individuals, trust, or appropriate community organisations—act as data stewards,[55] retaining greater control over how data is managed and shared. Data stewardship models improve

[54] Please see “Mozilla Data Collective,” Mozilla Foundation, https://datacollective.mozillafoundation.org.

[55] Data stewards act as custodians of the data, overseeing the data quality and use across its lifecycle, from creation to deletion. “What is Data Stewardship and Why is it Important,” Open Universal Science, 2024,https://opusproject.eu.

dataset discoverability and support responsible reuse by enforcing baseline documentation and quality standards. Creating metadata standards and ensuring that every dataset included in a national platform receives a unique identifier, such as a Digital Object Identifier (DOI), will help avoid duplication, improve discoverability, and support long-term tracking and reuse.[56]

India could pilot similar stewardship frameworks aimed at maintaining quality over time and sustaining hosting with community support. However, in their work on data trusts in the climate domain, researchers found that the feasibility of such models was lower in contexts like India, where digital literacy and infrastructure constraints persist.[57] This highlights the need for better capacity building and enabling infrastructure to employ alternative models. Piloting appropriate stewardship models will require long-term financing through blended-financing models, alongside capacity-building and engagement strategies to cultivate willing and qualified contributors capable of conducting regular quality checks against clearly agreed-upon community guidelines for data governance, and to foster community interest in managing their own datasets.

[56] GIZ, " A Study on Open Voice Data in Indian Languages," (Bonn, GIZ, 2020),https://www.bmz-digital.global.

[57] Vinay Narayan, "Data co-ops: Co-designing data trusts for climate action," 2023, https://thedataeconomylab.com.

5. Strengthening Responsible Deployment

5. Strengthening Responsible Deployment

Responsible deployment of speech technologies depends not only on building accurate models but also on ensuring that the systems introduced into public or commercial settings are safe, context-appropriate, and accountable. Several structural gaps currently undermine this goal. Before examining these gaps, it is essential to recognise that organisations face practical trade-offs when selecting models for deployment: balancing accuracy (model performance), cost (resource requirements), and reliability (consistency across diverse user populations).[58] These considerations become especially critical when structural gaps exist. These gaps manifest as mismatches across data practices and evaluation methods, foreseeable risks of misuse and unintended harm. Entities deploying AI systems often lack the resources to build their own foundational models. In such cases, the priority should be to identify existing models that align with the intended use case, then improve performance through targeted fine-tuning using more specific datasets, available tooling, and a focused data-collection effort where needed.

Organisations face practical trade-offs when selecting models for deployment: balancing accuracy (model performance), cost (resource requirements), and reliability (consistency across diverse user populations).

India’s voice technology ecosystem is at a promising inflection point, with a strong opportunity to build better institutions, safeguards, and governance frameworks to ensure it remains safe, fair, and inclusive for all. This section examines the challenges and learnings emerging from the Indian approach.

[58] Interview with Soma Dhavala, IIT Jammu, virtual, 24 September 2025.

5.1 Challenges

a. **Lack of Value Sharing:** Many speech datasets are built through the efforts of communities, researchers, and annotators who receive little recognition or return when their contributions fuel commercial innovation. Such practices occur across both crowdsourced data collection initiatives and secondary data acquisition through web scraping, although they operate via distinct mechanisms in each context. Addressing these challenges therefore requires thoughtful consideration beginning at the data-collection stage and continuing throughout the evolution of voice technologies.

Communities participating in data collection and curation for AI training often face a dilemma: keeping datasets open enables innovation and collaboration but allows well-resourced corporations to profit without giving back, while restricting access may protect against exploitation but undermine the collaborative benefits that make open-source AI powerful.[59] Some middle-ground solutions that can help communities balance openness with fairness include licensing conditions such as "share-alike" licenses, combined with capacity building and resource sharing.Further, in the absence of clear licensing terms, attribution requirements, or community-centred governance frameworks, open source speech datasets risk being absorbed into proprietary technologies without recognition or benefit-sharing with the communities, researchers, or organisations that contributed to them.[60] This concern is particularly acute in the Indian context, where the need for representative datasets drives data collection efforts in hitherto underserved areas, often led by small academic groups or grassroots initiatives operating on limited resources in the public interest. These contributors invest significant time and effort in data collection, annotation, and curation, often motivated by goals of linguistic preservation or equitable digital inclusion, yet rarely receive tangible returns on their investment.

Additional practical challenges related to attribution and value sharing arise when speech data is scraped from public platforms like YouTube, podcasts and social media[61]. Such datasets often lack the requisite metadata documenting speaker identity, demographic information,

[59] Open for Good Alliance, "Open-Source AI Data Sharing: yes! Data Colonialism: no!," October 2023, https://medium.com.
[60] Interview with Jagadish Babu, CEO, EkStep Foundation, virtual, 30 June 2025.
[61] Moaiad Ahmad Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application," International Journal of Advances in Soft Computing & Its Applications 13, no. 3 (2021), https://www.i-csrs.org.

and the recording context, undermining transparency in data provenance. This makes it difficult to assess how datasets were constructed, verify whether contributors understood or agreed to the potential use of their data, and establish responsibilities in the event of downstream consent change, misuse or harm.

When commercial developers or global corporations use this data without due credit or compensation, especially to build for-profit voice assistants, customer service bots, or surveillance tools, the practice becomes extractive. In many cases, the communities whose voices and languages power these systems are rarely informed, let alone included in decisions about how their data is used, monetised, or refined.[62] Moreover, when data is obtained from open source datasets curated with public finances, and there is no mechanism to track commercial usage, the absence of attribution further entrenches extractive data practices.[63] This creates a one-way flow of value, where local linguistic and cultural resources are transformed into commercial products that generate value elsewhere, with little or no return to the original contributors.In the absence of safeguards like open licenses with attribution clauses, copyleft clauses, data-use agreements, or benefit-sharing models, such exploitative practices risk deepening existing global inequities in AI development.

To address some of these concerns, the Indian government has recently proposed draft amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (Intermediary Rules) targeting synthetically generated information (SGI), or deepfakes. These amendments introduce due diligence obligations on certain intermediaries, requiring them to prominently label SGI using permanent, tamper-proof, unique metadata or identifiers. However,these measures raise significant practical challenges, as such labels and identifiers can potentially be circumvented or removed from speech datasets.

b. **Risks of Misuse:** Misuse refers to deliberate actions in which speech data is exploited beyond its original or intended purpose, often for profit, control, or deception. This includes malicious actors actively leveraging speech datasets to harm individuals or communities. Key types of misuse include:

- **Phishing and fraud:** Voice cloning and speech synthesis technologies enable malicious actors to convincingly impersonate

[62] Katie Shilton et al., “ExcavatingAwareness and Power in Data Science: A Manifesto for Trustworthy Pervasive Data Research,” Big Data & Society 8, no. 2 (2021): 20539517211040759, https://journals.sagepub.com.

[63] Interview with Jagadish Babu, CEO, EkStep Foundation, virtual, 30 June 2025.

trusted individuals—such as family members, government officials, or bank representatives—to carry out phishing scams or financial fraud. In the Indian context, several documented cases have already demonstrated how synthetic voice technologies are weaponised to manipulate individuals into transferring funds or disclosing sensitive information.[64] The ease with which many open source speech models can be re-engineered to perpetrate fraud, coupled with limited public awareness and established recourse mechanisms for such threats, significantly amplifies the risk of harm, particularly for digitally vulnerable individuals.

- **Misinformation and deepfakes:** Synthetic speech tools can be weaponised to fabricate public statements or interviews by political figures, producing audio deepfakes that are harder to detect than manipulated video.[65] Political parties have also reportedly deployed deepfakes and AI in aggressive pre-election campaigning, ranging from multi-lingual public addresses and personalised video messages to the creation of lifelike videos of deceased leaders.[66] In a country like India, with frequent elections and diverse linguistic and socio-cultural communities, voice-based deepfakes threaten democratic processes, trust in institutions, and public safety.

c. **Unintended Consequences:** Unlike misuse, unintended consequences emerge from design gaps, limited foresight, or systemic bias. They originate and accentuate across the data collection, curation and processing lifecycle, becoming apparent during downstream use. Some key consequences of use of voice technologies include:

- **Exclusion of marginalised groups:** Voice recognition systems trained exclusively on dominant accents or dialects can risk excluding communities whose speech patterns differ from the mainstream language. When such linguistically biased systems are integrated into public service delivery channels through information chatbots or grievance redressal systems, they risk excluding these communities from accessing public services. For instance, research shows that Hindi ASR models in India perform poorly for speakers from North-East and South India, highlighting the disproportionate focus on mainstream languages and dialects and the further marginalisation of low-resource languages.[67] In addition to exclusion based on language and dialects, ASR systems are known to exhibit

[64] Pankaj Sharma, “AI Scams Surge: Voice Cloning And Deepfake Threats Sweep India,” NDTV News, October 2024 https://www.ndtv.com.

[65] Namah Bose, “Deepfakes, Celebrities and Political Figures: Navigating the Rights of Celebrities and Political Figures in the Age of Deepfakes”, RGNUL Student Research Review, 27 July 27 2024, https://www.rsrr.in.

[66] Mitali Mukherjee, “AI Deepfakes, Bad Laws – and a Big Fat Indian Election”, Reuters Institute, March 2024, https://reutersinstitute.politics.ox.ac.uk.

[67] Javed et al., “Lahaja,” 2024.

performance disparity due to the gender, native region, age and speech rate of speakers.[68] Such disparities deepen digital inequities and limit access to services for marginalised populations.

- **Erosion of linguistic diversity:** The lack of support for regional dialects, accents, and low-resource languages also risks diminishing linguistic richness and cultural knowledge in model training and deployment.[69] When datasets disproportionately focus on dominant or "mainstream" languages such as Hindi, English, or other widely spoken state languages, they marginalise the vast spectrum of minority and tribal languages spoken by smaller or socioeconomically disadvantaged communities. This results in a case of "technolinguistic suspension", in which languages may be preserved in the cultural sense, but do not feature in digital spaces due to poor documentation or lack of access to AI tools. Such omissions pose challenges to the legitimacy and recognition of these languages in the digital age.[70] They are not merely technical gaps; they constitute a form of epistemic erasure, whereby specific ways of speaking, storytelling, and expressing identity are rendered invisible in digital systems.[71]

  Considerations of ease of access to better-resourced languages, efficiency interests, and assumptions about the dominance of certain languages can further erase key markers of regional, linguistic, and cultural identity to conform to dominant expectations.[72] For example, call centres use real-time accent neutralisation technologies to homogenise Indian accents in favour of Western accents.[73]

- **Fragmentation of institutional practices:** The absence of coordinated institutional practices for responsible deployment results in uneven standards for testing, assessment and accountability. Voice technologies are currently deployed across government departments, private companies, educational institutions, and civil society organisations, each with their own internal standards and governance mechanisms. For instance, some organisations use diversity wishlists and conduct rigorous bias audits, while others deploy systems with minimal validation and rely on general-purpose models. The lack of mandatory impact

[68] Anand Kumar Rai, Siddharth D. Jaiswal, and Animesh Mukherjee, "A Deep Dive into the Disparity of Word Error Rates across Thousands of NPTEL MOOC Videos," in Proceedings of the International AAAI Conference on Web and Social Media 18 (2024): 1302–14, https://ojs.aaai.org/index.php.

[69] Julia Barnett, "The Ethical Implications of Generative Audio Models: A Systematic Literature Review." In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (2023), 146-161, https://arxiv.org.

[70] M. Eriksson Krutrök and P. Poromaa Isling, "Technolinguistic Suspension and the Digital Futures of Minority Languages: AI and Meänkieli in Sweden," New Media & Society 0, no. 0 (2025),https://doi.org.

[71] Eriksson Krutrök and Poromaa Isling, "Technolinguistic Suspension."

[72] Ameena L Payne, Tasha Austin, and Aris M Clemons, "Beyond the Front Yard: The Dehumanizing Message of Accent-Altering Technology," Applied Linguistics 45 Issue 3, (June 2024): 553-60, https://doi.org.

[73] For example, see "About Us," Sanas, https://www.sanas.ai.

assessments for any use case further compounds these challenges. Consequently, voice technologies are deployed in sensitive contexts, including criminal justice, healthcare, and agriculture, without systematic risk assessment or mitigation.

5.2 Policy Recommendations

a. **Institutionalise Value-Sharing Mechanisms:** Models built using community-contributed data should incorporate shared-benefit structures. AI benefit sharing could include considerations for fair distribution of and access to opportunities and gains from AI . This can include redistribution of economic gains, technology transfer and local capacity-building programmes, co-authorship or attribution norms, non-proliferation of dangerous capabilities, clearly specifying the downstream use cases for the overall benefit of the community.[74] Value-sharing can strengthen trust, encourage participation, and counteract extractive data practices.

   Government-supported projects can encourage value-sharing through multiple mechanisms. As mentioned in Section 5.2(b) above, data stewardship models offer a powerful framework for institutionalising value-sharing mechanisms for community-contributed data. These models can operate on principles of community guardianship, using special licenses to ensure data is not sold and is used for the community's benefit, as demonstrated by Te Hiku Media in New Zealand.[75] Value can also be returned non-financially. For instance, in public-civic partnerships like Catalonia's Project Aina, citizens donate data to a digital commons in exchange for public benefits, such as language preservation and government-funding for local start-ups that develop apps using the Aina database in the Catalan language.[76] These stewardship models strengthen trust and counteract extractive practices by ensuring the value generated from community data flows back to its source.

   Data stewardship models offer a powerful framework for institutionalising value-sharing mechanisms for community-contributed data.

   Mandatory attribution of data principals in documentation, as well as acknowledgement of contributing communities and institutions could be considered, where possible. In addition, revenue-sharing

[74] Sumaya nur Adan et al, AI Benefit-Sharing Framework: Balancing Access and Safety (Oxford: AIGI, 2025), https://aigi.ox.ac.uk.
[75] Te Hiku Media, "Principles of Māori Data Sovereignty," 2018, https://static1.squarespace.com.
[76] Project Aina, https://projecteaina.cat.

mechanisms between individuals/communities who share data and commercial applications that use publicly funded datasets could be explored.

As is practised currently by organisations such as AI4Bharat, publicly funded datasets should be released under copyleft licensing that requires derivative works to remain open, preventing proprietary capture.

b. **Encourage Transparency Mechanisms:** Requiring data cards, model cards, risk assessments, and structured review processes before deployment can significantly improve transparency and accountability. Public-sector procurement guidelines can incorporate these requirements, requiring vendors to demonstrate responsible testing, documentation, and disclosure. Academic and civil society audits can further strengthen oversight.

   Government convening bodies, such as Bhashini, can consider establishing documentation standards for all voice technologies deployed in government services or procured using public funds. These requirements can be made contractually enforceable. Independent audits by academic institutions or civil society organisations can provide external validation.

c. **Encourage Contextualised Use Case–Appropriate Data Collection by Downstream Developers:** While foundational datasets offer the basis for generalisable application, additional layers of nuance are important for use-case specific applications. Such datasets should reflect the contextual realities of the target environment, including geographic, socio-linguistic, and demographic factors. However, this must occur within clear guardrails: explicit framing of the use case, clear identification of the beneficiary groups, robust consent processes, and validation by domain experts and community stakeholders. Institutional guidance can help teams determine the appropriate level of contextualisation without overfitting or excluding vulnerable groups.

   One of the recent mechanisms through which such an attempt has been made is through the publication of Language as Infrastructure - Field Guide for Inclusive Language AI in India, which contains reflections and practices for inclusive and responsible deployment of voice technologies in India that can be used by developers while designing their voice-based solutions.[77] While best practices like these are an essential starting point, they can eventually mature into formal

[77] Bhashini, Language as Infrastructure - Field Guide for Inclusive Language AI in India, 2025, https://bhashini.gov.in.

institutional guidance or standards for various steps in data collection, including defining the scope of an exercise, consent protocols, benchmarks, and performance thresholds. Bhashini can also consider sharing technical resources alongside guidelines to lower barriers for small organisations. This could include template consent forms in multiple languages and checklists for stakeholder involvement.

d. **Implement Clear Frameworks on Attribution of Responsibility:** In the context of harms arising from the risks of misuse, while existing laws sufficiently account for the various harms, there is presently no clear demarcation of responsibility on the relevant stakeholder in the AI value chain. This is an important policy consideration given that different participants in the AI ecosystem (e.g., foundation model developers, deployers, end-users, etc.) play distinct roles. Given this, it is essential to implement frameworks or provide legal guidance on attribution of responsibility to the relevant participant to ensure that a disproportionate burden is not imposed on the AI ecosystem overall. In this regard, the recommendation in the India AI Guidelines for amending the Information Technology Act, 2000, to clearly define the classification, obligations, and liability of actors in the AI value chain is a welcome move for fostering a more holistic and effective AI governance framework in India.

e. **Prevent Misuse through Safeguards, Accountability, and Public Literacy:** From a legal policy and governance standpoint, to mitigate the risks of misuse, the India AI Governance Guidelines recommend a combination of technical safeguards and legal reforms.

   The report also proposes establishing an expert committee to develop global standards for content authentication and recommends that a new AI Governance Group review the regulatory framework to tackle deepfakes through techno-legal solutions, alongside strengthening government enforcement capacity and public literacy. Measures such as watermarking and labelling using unique identifiers, while valuable in signaling the authenticity of the content, are not infallible measures.

   At the same time, citizen-facing literacy initiatives can also help users understand risks, exercise their rights, and identify misuse (e.g., voice phishing, deepfakes, unauthorised recording). Preventing harm requires both institutional enforcement and measures to empower users. When combined with strengthened government enforcement and greater public literacy, these measures would help combat the growing risks of AI misuse.

6. Conclusion: A Roadmap for Fair and Open Voice-Tech Innovation in India

6. Conclusion: A Roadmap for Fair and Open Voice-Tech Innovation in India

Voice technology is emerging as a key enabler for bridging linguistic gaps in India’s digital landscape. While the ecosystem remains nascent, significant progress has been made across the value chain, from data collection and curation to hosting and application development. However, current efforts remain fragmented, with actors working in parallel across proprietary and open-source domains and on both general-purpose and use-case-specific tools. These varied pathways reflect the diverse values and priorities of stakeholders, each approaching the opportunities and challenges of open voice technologies differently. For instance, while government, philanthropic, and academic interests lie in open-sourcing datasets and models to promote innovation and public use, downstream users may or may not value “openness” in isolation. For them, features like the ease of integrating the datasets and models for particular use cases or the cost of fine-tuning models for specific needs are essential considerations.[78] Coordination is therefore essential to ensure that these efforts are complementary and aligned with broader public-interest goals articulated by government stakeholders such as Bhashini.

In addition, public policy must move beyond narrowly correcting market gaps and instead actively shape the voice-technology ecosystem in line with broader social goals. India’s public-sector efforts, including Bhashini, already demonstrate this orientation toward inclusive digital transformation. Throughout this report, we have outlined practical recommendations that build on existing strengths while addressing institutional, technical, and governance gaps. These suggestions are intended to help align stakeholder incentives with long-term objectives of innovation and inclusion—whether by setting strategic direction, designing effective incentives, or building the institutional and infrastructural foundations needed for broad and equitable participation. Collectively, these measures offer a roadmap for government agencies, public institutions, and ecosystem partners to strengthen and sustain responsible voice-technology development.

[78] Interview with Varun Hemachandran, Lead Open NyAI, Agami, virtual, 14 July 2025.

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.

About 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.

About 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.

About 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.

About 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.

About 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)