ASR Datasets in Indian Languages: A Survey of Resources, Challenges, and Opportunities

KID: 20250306 | Ms Smitha HS

Abstract
Automatic Speech Recognition (ASR) has advanced rapidly in recent years, driven by deep learning models trained on large, high-quality datasets. While English ASR benefits from decades of investment and standardized corpora, Indian languages—with their linguistic diversity, dialectal variations, and code-switching practices—pose unique challenges. This article surveys publicly available and proprietary ASR datasets for Indian languages, examining their scale, diversity, quality, and applications. It highlights major initiatives such as AI4Bharat's IndicVoices, Kathbath, Shrutilipi, Dhwani, and Svarah, Mozilla Common Voice, OpenSLR (LDC-IL), Krutrim IndicST, and Bhasha Daan, alongside proprietary datasets from companies like Google, Microsoft, and Amazon. I contrast Indian ASR datasets with English resources, analyze the evolving ecosystem, and discuss implications for research, innovation, and inclusivity.

Introduction
Automatic Speech Recognition (ASR) relies on datasets that pair speech audio with text transcriptions. These datasets are critical for training models to convert spoken language into written form, enabling applications like virtual assistants, transcription services, and speech-to-speech translation. For India—a nation with 22 constitutionally recognized languages and hundreds of dialects—ASR datasets are essential for bridging digital divides and ensuring equitable access to speech technologies.

Despite India's linguistic richness, the development of ASR datasets faces challenges: scarcity of resources, noisy environments, dialect diversity, and widespread code-switching (e.g., Hinglish, Tanglish). This article surveys the landscape of Indian ASR datasets as of March 2025, highlighting both public and proprietary initiatives.

Anatomy of an ASR Dataset
An ASR dataset typically consists of:

- **Audio recordings** (formats like WAV/MP3, 8–48 kHz sample rates, mono/stereo).
- **Transcriptions** (manual, automatic, or phonetic).
- Annotations (speaker metadata, noise conditions, alignments).
- **Diversity attributes** (accents, domains, gender, age).
- **Licensing** (ranging from open-source CC0 to restricted proprietary).

High-quality datasets support model training, benchmarking, and deployment across diverse conditions.

Building ASR Datasets for Indian Languages
Constructing ASR datasets involves:

- **Data Collection** – Professional recordings, conversational speech, crowdsourcing, or mining online content.
- **Transcription** – Manual annotation, AI-assisted correction, or phonetic labeling.
- **Preprocessing** – Noise reduction, segmentation, normalization (e.g., "10 kg" → "ten kilograms").
- **Validation** – Word Error Rate (WER), human review, benchmarking.

Challenges for India include dialectal diversity, background noise, and multilingual mixing. Initiatives like IndicVoices, Shrutilipi, and Project Vaani aim to address these complexities.

Publicly Available Indian ASR Datasets:

AI4Bharat Initiatives

- **IndicVoices**: 12,000 hours (3,200 transcribed), 22 languages, highly diverse, CC-BY-4.0.
- **Kathbath**: 1,684 hours, 12 languages, professionally labeled, CC-BY-4.0.
- **Shrutilipi**: 6,400+ hours from All India Radio, 12 languages, broadcast speech, CC-BY-4.0.
- **Dhwani**: 17,000 hours unlabeled, 40 languages, mined from YouTube/News On AIR, MIT License.
- **Svarah & Lahaja**: Accent-focused benchmarks for Indian English and Hindi.

OpenSLR – LDC-IL
20–100 hours per language (Hindi, Tamil, Bengali, etc.), clean recordings, high-quality transcriptions, CC-BY-4.0.
Mozilla Common Voice
1,000+ hours for Hindi, smaller sets for Tamil, Telugu, Kannada, etc., crowdsourced, CC0 license.
Krutrim IndicST
10.8k hours training, 1.1k evaluation, 9 languages, curated from 14 open datasets plus synthetic data.
Bhasha Daan (Bhashini)
Citizen-contributed, growing repository covering 22 languages, variable quality, supporting India's National Language Tech Mission.
Proprietary Datasets
Several corporations maintain proprietary Indian language ASR datasets:

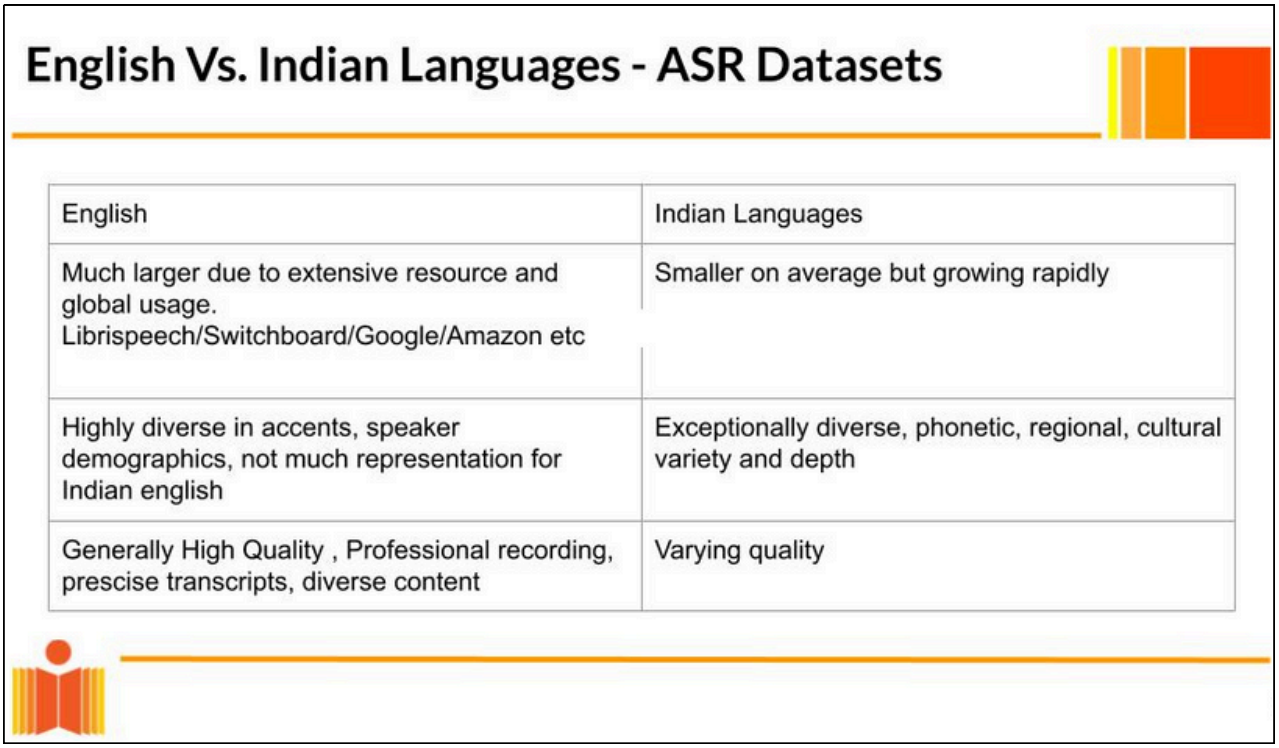
- **Microsoft Indian Language Corpus** – Telugu, Tamil, Gujarati, Hindi, Bengali (partially public, mostly proprietary).
- **Google Speech-to-Text Data** – Multilingual, large-scale, proprietary.
- **Amazon Alexa Data** – Hindi, Tamil, Telugu, Marathi, Indian English accents.
- **iFLYTEK** – Hindi, Bengali, Tamil.
- **Speech Ocean / Appen** – Commercially licensed datasets for Indian clients.

- **Startups (e.g., Reverie, Gnani.ai)** – Domain-specific proprietary collections.

These datasets are often larger and higher quality but restricted to internal use, raising concerns around transparency and inclusivity.

Comparison with English ASR Datasets

- **Size:** English datasets (e.g., LibriSpeech 1,000 hours, Switchboard 2,600 hours) are smaller than proprietary Indian corpora but generally cleaner and more mature. Indian datasets are growing (IndicVoices: 12k hours, Dhwani: 17k hours).
- **Diversity:** Indian datasets surpass English in linguistic and phonetic diversity due to 22+ languages and regional variations.
- **Quality:** English datasets have standardized, high-quality transcriptions; Indian datasets vary from clean (Kathbath) to raw (Dhwani).



The Emerging Ecosystem

The Indian ASR dataset ecosystem is complementary:

- **AI4Bharat** – Large-scale, diverse, open resources.
- **Common Voice & Bhasha Daan** – Crowdsourced, inclusive, democratizing.
- **OpenSLR (LDC-IL)** – Smaller but stable baselines.
- **Krutrim** – Hybrid datasets tailored for Speech LLMs.
- **Proprietary datasets** – Large-scale, high-quality, but inaccessible.

Together, they form a multi-pronged foundation for research and deployment.

Conclusion

ASR datasets in Indian languages have grown significantly in scale, diversity, and accessibility, bridging historical gaps with English corpora. Open initiatives like AI4Bharat and Bhasha Daan have democratized access, while proprietary datasets drive commercial deployment.

The path forward requires balancing inclusivity, quality, and sustainability—ensuring that India's linguistic complexity is represented fairly in digital systems.

Future efforts must focus on:

- Expanding underrepresented languages.
- Addressing code-switching and dialectal variance.
- Creating standardized benchmarks.
- Building public–private collaborations.

With these steps, Indian ASR datasets can empower not just research, but also inclusive human–machine interaction for one of the world's most linguistically diverse populations.

References:

- **IndicVoices:** Towards Building an Inclusive Multilingual Speech Dataset for Indian Languages - Sarvam AI, AI4Bharat (IIT Madras team) arXiv, March 4, 2024
- AI4Bharat, Krutrim AI Labs, HuggingFace, OpenSLR
- Indian Languages Corpus for Speech Recognition - SUSHIL VENKATESH KULKARNI AND SUKOMAL PAL(IEEE Conference Publication) - IEEE Xplore, March 19, 2020
- YouTube [https://www.youtube.com/watch?v=cmy2zf6CuH4&t , https://www.youtube.com/watch?v=n3YWy8fozAE https://www.youtube.com/watch?v=uMKe-oqsWHI
- https://arxiv.org

Ms Smitha HS
PG student
Dept of Heritage Science and Technology