Vaani

GET THE DATA

VAANI

Capturing the language landscape for an inclusive digital India

Sujith Pulikodan[1], Abhayjeet Singh[2], Agneedh Basu[1], Nihar Desai[1], Pavan Kumar J[1], Pranav D Bhat[1], Raghu Dharmaraju[1], Ritika Gupta[†], Sathvik Udupa[†], Saurabh Kumar[2], Sumit Sharma[†], Visruth Sanka[‡], Dinesh Tewari[3], Harsh Dhand[3], Amrita Kamat[*], Sukhwinder Singh[3], Shikhar Vashishth[3], Partha Talukdar[3], Raj Acharya[*], Prasanta Kumar Ghosh[2]

[1] ARTPARK, I-Hub @ IISc [2] Dept. of Electrical Engineering, IISc [3] Google DeepMind
[†] Work done while at IISc [‡] Work done while at ARTPARK [*] Work done while at Google DeepMind

Google ARTPARK AI & Robotics Technology Park, I-Hub @ IISc SPIRE Lab NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 · Track on Evaluations and Datasets

Why VAANI?

1,000+

mother tongues in India

grouped into 121 languages (Census 2011 [1]), yet voice technology remains limited to mainstream languages.

Existing open Indic speech corpora fall short in three ways:

1 · Not enough coverage

No open corpus is designed to cover India's non-scheduled languages at scale.

2 · Language-anchored sampling

A single "language" label hides regional, dialectal and social variation.

3 · Speech–text only

No visual grounding for multimodal models.

> **VAANI addresses all three:** it samples by district, not by language, and pairs every recording with the image that prompted it.

VAANI data summary

31,255 h
spontaneous speech
24.0M audio segments

2,043 h
manually transcribed
balanced across all districts

105
languages
4 language families

165
districts
28 states + 3 union territories

158K
speakers
52.8% female · ages 20–70

289K
image prompts
aligned image–speech–text triplets

Sampled by district, not by language

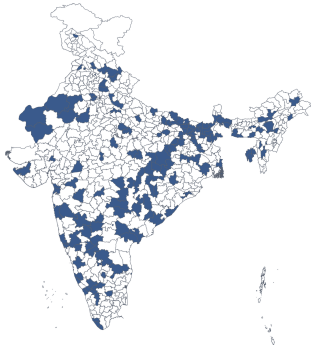


Figure: the 165 VAANI districts (blue) across 28 states and 3 union territories.

Districts were chosen to cover as many **under-represented and unrepresented languages** as possible across the country. Each district is recorded at several locations to capture variation within it.

PHASE 1

80 districts

12 states

PHASE 2

85 districts

23 states + 3 UTs

- ≈**200 h** of audio per district
- **5–10%** of each district manually transcribed
- ≥**800 speakers** per district, gender-balanced, ages 20–70

Spontaneous speech from images

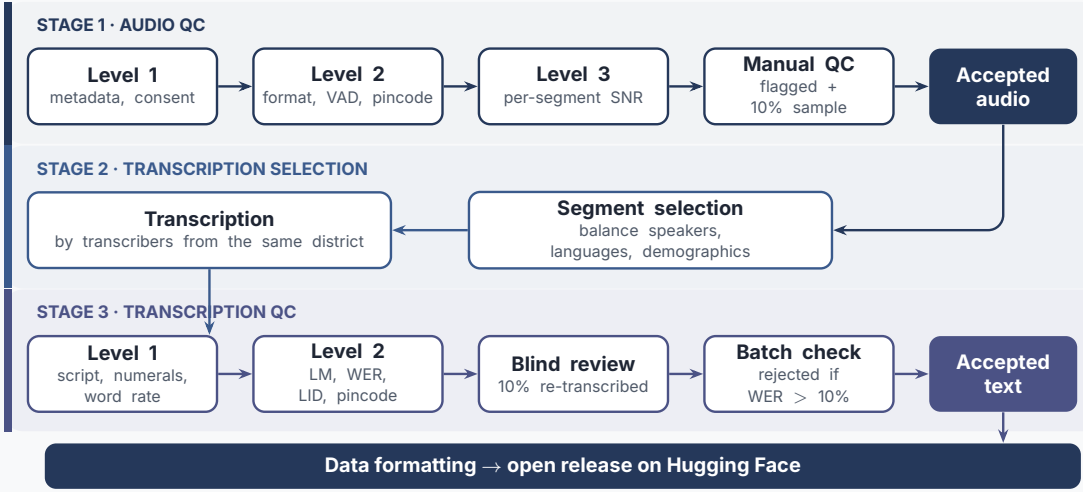




See an image → Speak 10–20 s → Transcribe

- **1,700–2,000 new photos per district**, local and generic, with no faces, logos or other PII.
- Speakers talk about each image in the **language they use at home**, so it works for non-literate speakers and unwritten languages.
- Up to 50 random images per session; speakers may skip any. Mean of **16.0 s** per image.
- Speakers are residents of the pincode where they are recorded; transcribers are **from the same district**, so they know the local dialect.

Multi-stage quality control

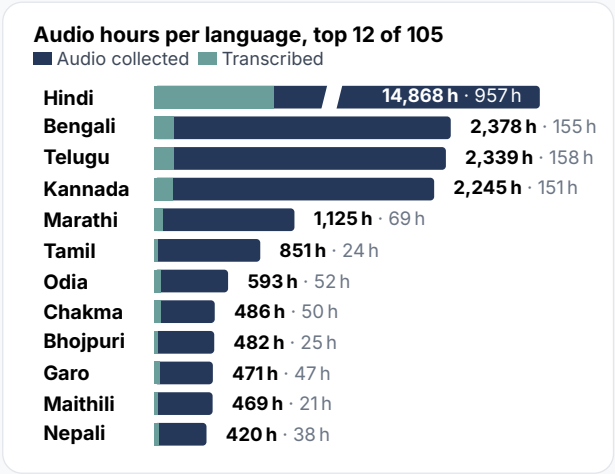


Compared with existing corpora

Dataset	Method	Audio (h)	Transcribed (h)	Languages	Speakers
VAANI	**Spontaneous**	**31,255**	**2,043**	**105**	**158K**
IndicVoices [2]	Mixed	23.7K	**11,200**	22	51K
RESPIN [3]	Read	10,416	10,416	9	18K+
Shrutilipi	Read	6,457	6,457	12	NA
SPRING-INX [4]	Mixed	5,271	5,271	14	NA
CommonVoice	Read	2,212	2,212	15	NA
Kathbath	Read	1,684	1,684	12	1,218
Gram Vaani	Spontaneous	1,108	108	1	NA
SYSPIN	Read	920	920	9	18
MUCS	Mixed	600	600	7	NA

VAANI and the 9 largest of 18 other Indic speech datasets by audio hours (paper Table 1). NA = not reported by the original authors. Read = recited text prompts · Spontaneous = unscripted, prompted by images or questions · Mixed = both.

105 languages, long tail included



+93 more languages

Many in an open speech dataset at this scale for the first time: Wancho · Idu Mishmi · Lepcha · Kokborok · Tulu · Halbi · Kurukh · Santali · Ao · Sumi · Paniya …

Four language families

Indo-European	Hindi, Bengali, Maithili, English
Dravidian	Telugu, Kannada, Tamil, Tulu
Sino-Tibetan	Garo, Mizo, Wancho, Lepcha
Austroasiatic	Santali

Fine-tuning on VAANI improves ASR

Whisper-large-v3-turbo [5]
79.3% → 35.9%

Parakeet-tdt-0.6b-v2 [6]
38.5% English-only, no Indic support

Gemma-3n-E2B (LoRA) [7]
44.4% → 33.2%

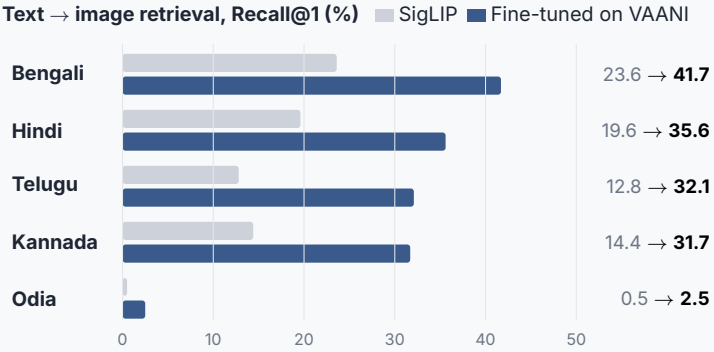
Pooled WER, zero-shot → fine-tuned, joint training on six languages. Parakeet-v2 is English-only, so it has no meaningful zero-shot baseline. Average WER on benchmarks by language:

	Whisper		Gemma-3n		Parakeet
Language	zero-shot	fine-tuned	zero-shot	fine-tuned	fine-tuned
Hindi	42%	**25%**	26%	**23%**	**26%**
Bengali	108%	**44%**	36%	**35%**	**47%**
Kannada	81%	**46%**	91%	**45%**	**51%**
Telugu	94%	**47%**	48%	**42%**	**51%**
Chakma	332%	**43%**	122%	**72%**	**49%**
Bhojpuri	73%	**25%**	58%	**36%**	**28%**

Significant in all six languages ($p < 0.002$, paired bootstrap + Wilcoxon, Holm-corrected; 65,320 test utterances). Splits are disjoint by speaker. Chakma and Bhojpuri tested on VAANI only.

Text-to-image retrieval

For 10,287 images, all spoken descriptions of each image were combined into one per language (14,564 in total), and multilingual SigLIP [8] was fine-tuned on these image–description pairs.



Pooled Recall@1
10.6% → 27.3%

Median rank
29 → 4

Over 1,045 test images. Odia improved least, possibly because of tokenizer fragmentation.

Limitations and future work

Limitations

- Only **2,043 of 31,255 h (∼6.5%)** are transcribed; we prioritised balanced transcription across districts over depth.
- Coverage is small for many of the 105 languages; most data is in a few majority languages.
- **165 of ∼800 districts** so far, so many dialects and minority languages remain unrepresented.
- Spontaneity of speech was not verified per utterance.
- ASR results are aggregate, not broken down by age, gender, education or socio-economic status.

Future work

- Expand VAANI to **additional districts** in future phases.
- Improve **audio and transcription coverage** for low-resource languages.
- Use the demographic metadata for **fairness analyses** across age, gender, education and socio-economic status.

Conclusion

01

A district-sampled, multimodal resource

31,255 h of audio, 2,043 h of transcriptions and 289,838 images, with rich metadata across 165 districts and 105 languages.

02

Useful today

Fine-tuning on VAANI lowers ASR error and improves text-to-image retrieval, supporting models that generalise beyond standard dialects.

03

Open for everyone

Released under CC-BY-4.0 to lower the barrier to Indic speech research, with future phases adding districts and low-resource transcription.

VAANI is a foundational resource for voice technologies that meaningfully serve under-represented linguistic communities across India.

Acknowledgements

Funding

We gratefully acknowledge **Google** for funding this work.

Support & guidance

The leadership of **Google**, the **Indian Institute of Science (IISc)** and **ARTPARK**, throughout Project Vaani.

Participants

The people across India who contributed their voices: about **158K speakers** in **165 districts**.

Field & data teams

The **data collection vendors** who made large-scale field collection possible, and the **data curators and transcribers** whose careful work made the dataset usable.

References

[1] Government of India. Census of India 2011: Languages. *2011.*

[2] T. Javed et al. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. *Findings of ACL, 2024.*

[3] S. Kumar et al. RESPIN-S1.0: A read speech corpus of 10000+ hours in dialects of nine Indian languages. *NeurIPS Datasets & Benchmarks, 2025.*

[4] A. Gangwar et al. SPRING-INX: A multilingual Indian language speech corpus by SPRING Lab, IIT Madras. *arXiv:2310.14654, 2023.*

[5] A. Radford et al. Robust speech recognition via large-scale weak supervision (Whisper). *ICML, 2023.*

[6] NVIDIA. parakeet-tdt-0.6b-v2. *Hugging Face, 2025.*

[7] Gemma Team. Gemma 3n. *Google DeepMind, 2025.*

[8] M. Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786, 2025.*

Thank you

Questions?

Write to us at **vaanicontact@gmail.com**

GET THE DATA

huggingface.co/datasets/ARTPARK-IISc/Vaani · CC-BY-4.0

vaani.iisc.ac.in

Google ARTPARK AI & Robotics Technology Park, I-Hub @ IISc SPIRE Lab