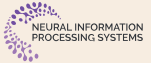
# RESPIN-S1.0: A read speech corpus of 10000+ hours in dialects of nine Indian Languages

Saurabh Kumar[1], Abhayjeet Singh[1], Deekshitha G[1], Amartyaveer[1], Jesuraj Bandekar[1], Savitha Murthy[1], Sumit Sharma[1], Sandhya Badiger[2], Sathvik Udupa[3], Amala Nagireddi[1], Srinivasa Raghavan K M[4], Rohan Saxena[4], Jai Nanavati[4], Raoul Nanavati[4], Janani Sridharan[4], Arjun Mehta[4], Ashish Khuraishi K S[4], Sai Praneeth Reddy Mora[4], Prashanthi Venkataramakrishnan[4], Gauri Date[4], Karthika P[4], Prasanta Kumar Ghosh[1]

[1]Dept. of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India
[2]German Research Center for Artificial Intelligence (DFKI), Germany
[3]BUT Speech@FIT, Brno University of Technology, Czech Republic
[4]Navana Tech, Mumbai, India

spirelab.ee@iisc.ac.in prasantg@iisc.ac.in


NeurIPS 2025, Datasets & Benchmarks Track

# Overview

# Overview

# Motivation: Dialectal Gap in Indian ASR

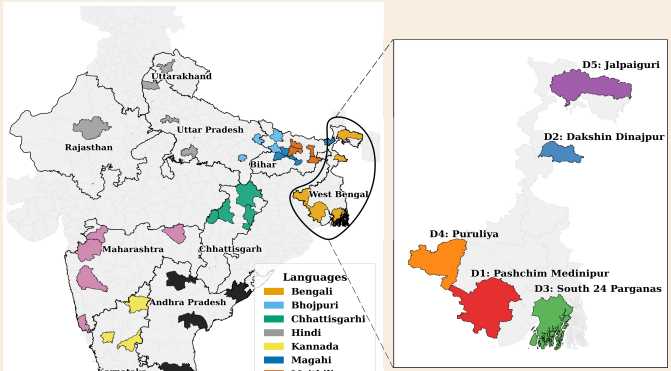
- Indian speech tech often trained on **standard/urban** varieties → poor performance for rural, low-literacy speakers.
- Public-service domains (agriculture, banking/finance) are especially affected.
- **Objective:** release a large, dialect-rich, ethically collected read-speech corpus to support ASR and dialect identification (DID) research.

**Why RESPIN-S1.0?**

- Source-level dialect representation.
- Domain-grounded text.
- Reproducible splits for benchmarks.

# Overview

# Geographic and Dialectal Coverage




*Geographic distribution of RESPIN-S1.0 dialect coverage across India with a zoom-in on*

## Overview

- Covers **9 major Indian languages** and **38+ district-linked dialects**.
- Dialect regions explicitly mapped to **Census 2011 administrative districts**, enabling reproducible and geographically grounded analysis.
- Facilitates comparison across **Indo-Aryan** and **Dravidian** linguistic families within a unified corpus.

# RESPIN-S1.0: Summary

- **Languages:** 9 (bn, bh, ch, hi, kn, mg, mt, mr, te)
- **Dialects:** 38+ district-linked varieties
- **Speech:** 10,416 h Clean + 2,288 h Semi-noisy + 1,617 h Noisy = **14,322 h**
- **Speakers:** ∼18k from 1,500+ pincodes
- **Domains:** Agriculture & Banking/Finance
- **Text:** 209k+ native-authored sentences
- **Ethics:** IISc approval, mobile-app consent, fair compensation

# Overview

# Text Corpus Design

- ∼2.1L native-authored prompts across the two target domains.
- 3–5 dialects per language; domain balance enforced at authoring time.
- ∼3.6% sentences filtered during validation for mismatch/anomalies.
- Limited translation (6–17%) only where dialectal gaps existed; Bengali is fully native-authored.
- **Also released:** dialect-specific phonetic lexicons aligned with validated text (see paper/appendix for details).

# Lexicon and Text Statistics by Language

| Lang | Lexicon | | | | Text Corpus | | | |
|---|---|---|---|---|---|---|---|---|
| | Chars | Phones | LexWds | #Dials | #Sents | TotalWds | Vocab | AvgLen |
| bh | 71 | 54 | 14.1k | 3 | 23,916 | 250,542 | 14,000 | 3.75 |
| bn | 64 | 50 | 18.6k | 5 | 21,959 | 213,648 | 18,631 | 4.73 |
| ch | 68 | 50 | 13.2k | 4 | 21,281 | 317,907 | 13,184 | 3.36 |
| hi | 72 | 55 | 16.6k | 5 | 21,782 | 255,543 | 16,460 | 3.94 |
| kn | 66 | 50 | 50.8k | 5 | 25,178 | 235,511 | 50,622 | 6.51 |
| mg | 72 | 54 | 21.7k | 4 | 23,281 | 253,834 | 21,600 | 3.87 |
| mr | 68 | 51 | 35.7k | 4 | 25,012 | 220,768 | 35,464 | 5.47 |
| mt | 72 | 55 | 19.3k | 4 | 25,803 | 286,950 | 19,197 | 3.87 |
| te | 63 | 48 | 39.2k | 4 | 21,772 | 191,238 | 39,015 | 6.42 |

Number of dialects (#D), sentences (#S), total words (TotW), vocabulary size (Vocab), and average sentence length (AvgL).

**Observations**

- Dravidian languages (kn, te) → highest lexical diversity and longer sentences.
- Indo-Aryan languages (bh, ch, mg) → more compact vocabularies.

# Audio Statistics and Quality Slab Distribution

| Lang | Slab-wise Duration (h) | | | Aggregate Audio Statistics | | | |
|---|---|---|---|---|---|---|---|
| | Clean | Semi | Noisy | Total(h) | #Utts | #Spkrs | AvgDur(s) |
| bh | 1116.96 | 105.87 | 76.57 | 1299.4 | 982,394 | 1,532 | 4.76 |
| bn | 1213.24 | 189.96 | 6.73 | 1409.9 | 986,295 | 1,964 | 5.15 |
| ch | 1295.97 | 198.27 | 90.12 | 1585.0 | 966,754 | 1,777 | 5.90 |
| hi | 944.59 | 362.79 | 390.86 | 1698.6 | 1,231,540 | 2,525 | 4.97 |
| kn | 1344.10 | 250.10 | 226.23 | 1671.0 | 990,756 | 2,009 | 6.07 |
| mg | 1279.48 | 112.39 | 72.74 | 1529.5 | 1,077,792 | 2,239 | 5.11 |
| mr | 1379.44 | 335.21 | 199.94 | 1814.6 | 1,353,012 | 2,391 | 4.83 |
| mt | 723.06 | 510.24 | 405.29 | 1638.6 | 1,048,311 | 2,076 | 5.63 |
| te | 1303.98 | 222.53 | 148.99 | 1675.5 | 1,114,563 | 2,064 | 5.41 |
| **Total** | **10,416** | **2,288** | **1,617** | **14,322** | **9,751,417** | **18,577** | – |

**Key Insights**

- **14.3k h** validated across 9 languages and 18k+ speakers.
- **Clean slab (>10k h)**: benchmark-quality subset for ASR and DID.
- **Semi/Noisy slabs**: support robustness, domain-shift, and fairness studies.
- **Short avg. durations (4–6 s)**: suitable for fine-grained alignment and acoustic modeling.

# Validation Process and Slab Assignment

## Multi-stage Validation Pipeline

- **Automatic Scoring:** Forced alignment and semi-automated models compute alignment confidence scores for every utterance.
- **Manual Verification:** $\sim$5% of utterances per dialect manually inspected to calibrate and refine thresholds.
- **Slab Assignment:** Recordings categorized into **Clean**, **Semi-noisy**, and **Noisy** based on scoring outcomes, verified through multi-stage manual validation.

## Design Rationale

- Ensures a high-quality benchmark core while preserving naturally occurring variability.
- Facilitates reproducible evaluation across clean and challenging conditions.
- Provides a transparent and scalable quality-control pipeline for future RESPIN releases.

# Corpus Design, Validation, and Ethical Safeguards

## Dialectal and Data Integrity

- All stages—text creation, speaker recruitment, and validation—conducted *per dialect* to preserve linguistic authenticity.
- ∼1500 domain-specific subtopics (e.g., soil health, crop insurance, UPI, KYC) curated for balanced domain coverage.
- Dialect-aware reviewer workflow ensured consistent text–audio–dialect alignment across composing, recording, and validation stages.

## Ethical and Demographic Safeguards

- Speaker enrollment via the **Bolo mobile app** with WhatsApp-based verification and informed consent.
- Participant demographics (age, gender, region) verified for internal consistency before release.
- **Ethics approval** obtained from IISc; contributors received **fair compensation** in compliance with research standards.

# Overview

# Benchmarking Setup

**Models evaluated**

- TDNN-HMM [2]
- E-Branchformer [3]
- Whisper (Tiny/Base/Small) [4]
- IndicWav2Vec, SPRING-Wav2Vec2, SPRING-Data2Vec-AQC

**Protocol**

- All trained/fine-tuned only on RESPIN standardized train splits.
- Evaluated per language using common dev/test.

# Benchmarking on RESPIN-S1.0

| Model | Avg. CER (%) | Avg. WER (%) |
|---|---|---|
| *Pretrained (non-RESPIN fine-tuning)* | | |
| SeamlessM4T-v2-Large | 22.23 | 51.81 |
| IndicW2V | 17.09 | 49.25 |
| SPRING-W2V2 | 13.29 | 39.58 |
| SPRING-Data2Vec-AQC | 12.88 | 38.83 |
| *Traditional (RESPIN scratch)* | | |
| TDNN-HMM | 4.99 | 17.28 |
| E-Branchformer | 4.52 | 16.63 |
| *Fine-tuned on RESPIN* | | |
| Whisper-Tiny | 10.66 | 32.44 |
| Whisper-Base | 7.31 | 25.36 |
| Whisper-Small | 5.69 | 20.08 |
| IndicW2V | 4.60 | 19.84 |
| SPRING-W2V2 | 3.71 | 15.92 |
| **SPRING-Data2Vec-AQC** | **3.58** | **15.40** |

*Averages over 9 RESPIN languages.*

**Key empirical findings**

- **Best overall:** SPRING-Data2Vec-AQC (RESPIN fine-tuned) achieves the lowest average error: **3.58% CER**, **15.40% WER**.
- **Effect of RESPIN fine-tuning:** Whisper and SSL models substantially improve over their non-RESPIN counterparts → dialectal adaptation is essential.
- **Pretrained (non-RESPIN) models** underperform on dialect-rich speech (WER ∼39–52%), showing domain mismatch.
- **Traditional RESPIN-scratch models** (TDNN-HMM, E-Branchformer) remain competitive baselines (WER 16–17%) for lower-capacity settings.
- **Overall:** RESPIN-S1.0 enables a clear ranking of ASR systems under Indian dialectal variation.

# Applications, Impact, and Limitations

**Applications**

- MADASR 2023/2025 and other dialect-aware ASR challenges.
- ASR, DID, LID, robustness, fairness evaluations.

**Impact**

- Focus on agriculture & finance → directly useful for rural public-service speech systems.
- Reproducible, open splits meet NeurIPS D&B expectations.

**Limitations**

- Read speech only; spontaneous/conversational planned.
- Domains limited to two verticals; future: healthcare, education, governance.
- Smartphone-based collection may miss non-digital users.

# Overview

# Conclusion and Future Directions

**Conclusion**

- RESPIN-S1.0: 9 languages, 38+ dialects, 14k+ hours, dialect-preserving pipeline.
- Ships with text, phonetic lexicons, metadata, and benchmark-ready splits.

**Future work**

- Extend to spontaneous / conversational speech.
- Add more public-service domains.
- Release companion benchmark recipes for SSL and large ASR models.

# References I

Office of the Registrar General & Census Commissioner, India, "Census of India 2011." https://censusindia.gov.in/census.website/, 2011. Accessed on 6 November 2025.

D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech 2018*, pp. 3743–3747, 2018.

K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-Branchformer: Branchformer with Enhanced Merging for Speech Recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 84–91, 2023.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org, 2023.

# Acknowledgements

**Project Support**

- This work was supported by the **Gates Foundation**.
- Conducted at the **SPIRE Lab**, Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore.

**Collaborating Partners**

- We thank our partners **Navana Tech** and **ARTPARK**, along with NGOs, volunteers, and community contributors, for their dedicated efforts in data collection, validation, and annotation.
- Their contributions were instrumental in developing the large-scale, dialect-rich **RESPIN-S1.0** corpus.

# Thank You!

## Questions or Suggestions?

*Scan for paper, dataset, and resources*

**Code Repository**
github.com/labspire/respin_baselines

**OpenReview Link**
openreview.net/forum?id=qL8M2dOY4L

**Contact**
saurabhk0317@gmail.com
spirelab.ee@iisc.ac.in
prasantg@iisc.ac.in