MADASR 2.0: Multi-Lingual Multi-Dialect ASR Challenge in 8 Indian Languages

Saurabh Kumar[1], Sumit Sharma[1], Deekshitha G[1], Abhayjeet Singh[1], Amartyaveer[1], Sathvik Udupa[2], Sandhya Badiger[3], Sanjeev Khudanpur[4], Sunayana Sitaram[5], S. Umesh[6], Bhuvana Ramabhadran[7], Brian Kingsbury[8], Hema A. Murthy[6], Srikanth S. Narayanan[9], Howard Lakougna[10], Prasanta Kumar Ghosh[1]

[1]Indian Institute of Science (IISc), Bengaluru, India [2]BUT Speech@FIT, Brno University of Technology, Czech Republic [3]German Research Center for Artificial Intelligence (DFKI), Germany
[4]Johns Hopkins University, USA [5]Microsoft Research, India [6]Indian Institute of Technology Madras (IITM), India [7]Google DeepMind, USA [8]IBM Research, USA
[9]University of Southern California (USC), USA [10]Gates Foundation, USA

Introduction & Motivation

Motivation

- India has 100+ languages and many dialects with substantial phonological variation, making ASR difficult.
- SSL models (wav2vec 2.0 [1]) and multilingual ASR (Whisper [2]) still underperform on low-resource Indian languages.
- Practical ASR must handle dialect shifts, mixed registers, and read–spontaneous mismatch.
- MADASR 1.0 (ASRU 2023) addressed Bengali & Bhojpuri dialect ASR; MADASR 2.0 expands to a multilingual, multidialect setting.

Scope of MADASR 2.0

- Includes **8 languages** and **33 dialects** using a curated RESPIN subset [3].
- Provides a benchmark for dialect-robust, cross-domain multilingual ASR.

Challenge Design

- Train on read speech; evaluate on read & spontaneous sets.
- Four tracks based on resource level & external data:
  - **1:** Constrained, 30h/lang
  - **2:** Constrained, 150h/lang
  - **3:** Unconstrained, 30h/lang
  - **4:** Unconstrained, 150h/lang
- Tasks: ASR (primary); LID & DID (optional).

Dataset: RESPIN Subset

Languages & Dialects

- 8 languages: Bengali, Bhojpuri, Chhattisgarhi, Kannada, Magahi, Maithili, Marathi, Telugu.
- District-level dialect labels yield **33 dialects**.

Data Splits

Split	Hours	Utterances	Speakers
Train-small	241.1	146,760	9,387
Train-large	1,239.5	733,800	11,315
Dev-read	17.6	11,507	660
Test-read	26.7	17,550	1320
Test-spont	7.0	5,400	–

Transcriptions

- Each utterance includes LID, DID, speaker ID, and transcript.
- Native scripts retained; English items written in textual/native-script form.
- No additional normalisation; dialect variation preserved.

Challenge Setup & Evaluation

Submission Portal

- React-based portal for team registration and password-protected submissions.
- Backend parses TSV outputs, scores automatically, and updates leaderboards.

Evaluation Protocol

- ASR scored via CER/WER using `jiwer`; invalid utterances removed.
- Punctuation removed; no further filtering or normalisation.
- LID & DID evaluated via exact token match.

Submission Format: ASR only; ASR+LID (e.g., [bn]); or ASR+LID+DID (e.g., [bn_D3]).

Baseline Systems & Key Findings

ESPnet Baselines

- Hybrid CTC/Attention ASR using a Conformer encoder and Transformer decoder.
- Tracks 1–2: supervised training on 30h/150h read-speech subsets.
- Tracks 3–4: same architecture with encoder initialised from frozen IndicWav2Vec SSL features.

Dev-Set Performance

Track	CER	WER	LID Acc.
1 (30h)	4.06	17.28	97.41
2 (150h)	**3.61**	**15.72**	96.58
3 (30h+SSL)	4.36	18.45	96.92
4 (150h+SSL)	3.89	16.74	96.18

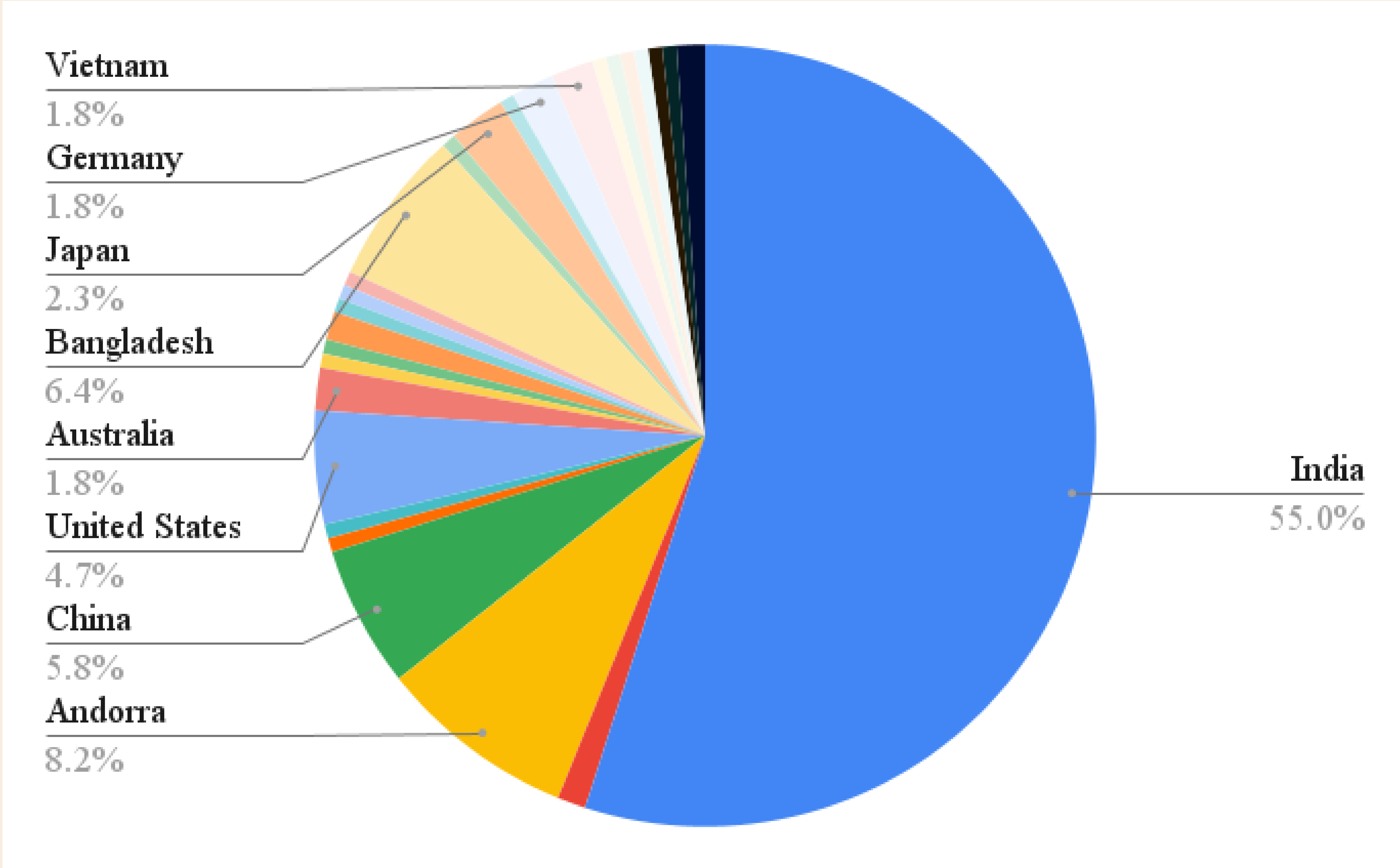
Observations

- More supervised data consistently improves CER/WER.
- SSL initialisation is competitive but not consistently superior to supervised training.
- Kannada and Telugu exhibit higher error rates, particularly for spontaneous speech.

Participation & Submitted Systems

Participation Overview

- 171 dataset downloads from 28 countries.
- 80 registered teams across 8 countries.
- 306 submissions received; 191 valid for ranking.
- 7 teams submitted at least one valid final system.



Geographic distribution of participants

Representative Submitted Systems

- **Team YS:** Whisper-large-v3 with MixLoRA for parameter-efficient finetuning.
- **SPRING Lab:** Common Label Set (via G2P) + dual-decoder architecture for LID/DID.
- **QWER:** voice conversion, mixed-speech augmentation, error-aware training.
- **Pramiteeh:** multitask ASR–LID–DID model with gender and age prediction.

Leaderboard Results

Best Non-baseline Systems Across Tracks

Track	Team	CER	WER	LID / DID
Read Speech				
1	QWER	4.84	18.68	12.73 / NA
2	SPRING Lab	4.40	17.06	97.39 / 75.36
3	YS	4.98	18.83	96.54 / NA
4	pramiteeh	6.71	24.53	96.00 / 62.58
Spontaneous Speech				
1	pramiteeh	23.67	59.80	80.02 / NA
2	pramiteeh	24.50	61.00	77.61 / 26.52
3	pramiteeh	23.66	59.79	80.02 / NA
4	pramiteeh	24.50	61.00	77.61 / 26.52

Highlights

- Read-speech CER ranges **4.30–8.07%**; Track 2 delivers strongest performance.
- SPRING Lab achieves best DID accuracy (**75.36%**) and strong LID.
- Spontaneous-speech CER degrades to **23.66–67.52%** across systems.
- pramiteeh consistently leads on spontaneous sets across all tracks.

Key Takeaways

ASR Performance & Data Effects

- Track 2 (constrained, 150h) delivers the best overall ASR results.
- Larger in-domain training data clearly improves model performance.
- Spontaneous speech causes major CER/WER degradation across all systems.
- Domain and style mismatch remain a key challenge.

Modelling Approaches

- External resources in Tracks 3–4 show mixed impact—some systems benefit, others suffer due to poor domain adaptation.
- Multitask learning (Pramiteeh) improves LID/DID accuracy.
- Script-agnostic CLS-based modelling (SPRING Lab) also boosts LID/DID without harming ASR.

Conclusion & Outlook

Summary

- MADASR 2.0 benchmarks multilingual, multidialect ASR (8 languages, 33 dialects).
- Read & spontaneous sets enable controlled robustness evaluation.

Future Directions

- Better spontaneous-speech & disfluency modelling.
- Stronger end-to-end dialect-aware ASR with LID/DID conditioning.
- Domain adaptation for dialect/generalisation gaps.

Acknowledgements Supported by the RESPIN project (Gates Foundation).

Links & Resources

- Challenge: `sites.google.com/view/respinasrchallenge2025`
- Corpus: `spiredatasets.ee.iisc.ac.in/respincorpus`
- Baseline: `github.com/saurabhk0317/espnet_respin_asru25.git` (branch: `respin_asru25`, path: `egs2/respin_asru25/asr1`)

Scan for paper

References

[1] A. Baevski *et al.*, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] A. Radford *et al.*, *Robust Speech Recognition via Large-Scale Weak Supervision*, 2022. arXiv: 2212.04356 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.04356.

[3] S. Kumar *et al.*, "RESPIN-S1.0: A read speech corpus of 10000+ hours in dialects of nine Indian Languages," in *Proc. NeurIPS, Datasets and Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=qL8M2dOY4L.

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