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[1], Sandhya Badiger[1], Sanjeev Khudanpur[2], Sunayana Sitaram[3], S. Umesh[4], Bhuvana Ramabhadran[5], Brian Kingsbury[6], Hema A. Murthy[4], Srikanth S. Narayanan[7], Howard Lakougna[8], Prasanta Kumar Ghosh[1]

[1]Indian Institute of Science (IISc), Bengaluru, India [2]Johns Hopkins University, USA [3]Microsoft Research, India
[4]Indian Institute of Technology Madras (IITM), India [5]Google DeepMind, USA [6]IBM Research, USA
[7]University of Southern California (USC), USA [8]Gates Foundation, USA

***Abstract*—We present MADASR 2.0, a challenge at ASRU 2025 aimed at advancing multilingual and multidialectal automatic speech recognition (ASR) in low-resource Indian languages. Building on the 2023 edition, it introduces a subset of the RESPIN corpus, over 1200 hours of read speech across 8 languages and 33 dialects, with test sets including both read and spontaneous speech. The challenge comprises four tracks varying by training data size and external resource usage, and supports auxiliary tasks like language and dialect identification. We detail the dataset, tasks, baselines, and submissions and analyse trends across tracks and speech styles. Results highlight the continued difficulty of spontaneous ASR, the benefits of multitask and transfer learning, and effective strategies for building dialect-aware ASR systems. MADASR 2.0 offers a standardised benchmark to support future research on inclusive and scalable ASR for linguistically diverse populations.**

***Index Terms*—Multilingual ASR, Dialectal ASR, Indian Languages, Low-Resource Speech Recognition, Benchmark Challenge, RESPIN Corpus**

I. Introduction

Recent advances in automatic speech recognition (ASR) have been fueled by self-supervised learning (SSL) methods such as wav2vec 2.0 [1] and large-scale multilingual models like Whisper [2] and Massively Multilingual ASR [3]. Despite these developments, robust ASR for low-resource Indian languages remains a significant challenge [4], [5]. India is home to over 100 spoken languages, including 22 constitutionally recognised ones, and hundreds of dialects that differ widely in phonology, lexicon, and prosody [6], [7]. These linguistic variations, often accompanied by limited annotated resources, hinder the development of scalable and inclusive ASR systems. Furthermore, speakers frequently alternate between standard and regional forms, especially in informal and multilingual settings, adding to the complexity of real-world ASR deployment [8]–[11].

While multilingual ASR systems [11]–[17] have made notable global progress, relatively little attention has been given to the intra-language dialectal diversity within Indian languages, particularly under realistic resource constraints. Recent research has demonstrated the promise of techniques like transfer learning, multimodal feature fusion, and multitask modelling in addressing dialectal variation [18], [19]. However, the lack of publicly available benchmarks and standardised evaluation protocols continues to limit progress.

To address this gap, the Model Adaptation for ASR (MADASR) challenge series was initiated. The first edition, MADASR [20], was held at ASRU 2023 and focused on monolingual dialectal ASR for Bengali and Bhojpuri [21]–[23]. It provided a curated read-speech corpus to evaluate adaptation strategies across dialects of a single language and highlighted the challenges posed by dialectal variation even within controlled data settings.

Building on this foundation, MADASR 2.0, organised at ASRU 2025, significantly expands both the linguistic and experimental scope. The challenge uses a curated subset of the RESPIN corpus [24], comprising approximately 1200 hours of transcribed speech spanning 8 low-resource Indian languages, Bengali, Bhojpuri, Chhattisgarhi, Kannada, Magahi, Maithili, Marathi and Telugu, comprising 33 dialects [18]. It includes only *read* speech in the training set, while both *read* and *spontaneous* speech test sets are provided for evaluation, thereby capturing greater acoustic and linguistic variability across domains, styles and speakers.

Unlike MUCS 2021 [7], which focused primarily on multilingual ASR and code-switching, MADASR 2.0 emphasises dialectal robustness and generalisation. It features four tracks, varying by the size of training data and the use of external resources, supporting both constrained and unconstrained development paradigms. The challenge further encourages multitask learning by allowing optional submissions for language identification (LID) and dialect identification (DID), facilitating a more comprehensive evaluation of multilingual and multidialectal models.

Separate leaderboards are maintained for read and spontaneous test sets, acknowledging the differences in linguistic complexity and acoustic conditions between the two styles. The RESPIN dataset used in this challenge has already supported prior research on joint ASR-DID modelling with multimodal features [18], [19], and MADASR 2.0 aims to standardise evaluation in this space and stimulate further advances.

By releasing a dialect-rich benchmark, clear evaluation protocols, and a public leaderboard [25], MADASR 2.0 aspires to catalyse progress in developing robust, equitable, and generalizable ASR systems tailored for linguistically diverse populations.

II. Challenge Design

A. Track Overview

To support a diverse set of research goals and ensure fair benchmarking, the MADASR 2.0 Challenge was divided into four tracks that vary in training resource availability and the allowance of external data. All tracks use subsets of the RESPIN corpus. Tracks 1 and 3 include 30 hours of training data per language (approximately 240 h total), while Tracks 2 and 4 use the full 150 h per language (approximately 1200 h total). The constrained tracks (1 and 2) restrict training to RESPIN data only, whereas the unconstrained tracks (3 and 4) permit the use of publicly available corpora and pretrained models.

- **Track 1:** Constrained low-resource track using 30 h per language. Only RESPIN data allowed.
- **Track 2:** Constrained high-resource track using 150 h per language. Only RESPIN data allowed.
- **Track 3:** Unconstrained low-resource track with 30 h per language, plus any external corpora or pretrained models.
- **Track 4:** Unconstrained high-resource track with 150 h per language and full use of external data.

Participants had access to read speech recordings and corresponding transcripts for training. These recordings were captured in natural, uncontrolled environments using prompted sentences spoken aloud by individual speakers. For testing, both read and spontaneous speech were included. The spontaneous speech set comprised conversational audio between two speakers on everyday topics, also recorded and transcribed in real-world conditions. This allowed the evaluation of model generalisation to out-of-domain speech.

B. Tasks and Leaderboards

The primary task across all tracks was automatic speech recognition (ASR). System performance was evaluated using character error rate (CER) and word error rate (WER) on both read and spontaneous test sets. Each track thus featured two leaderboards, one for read speech and one for spontaneous speech, to capture both in-domain and cross-domain accuracy.

In addition to ASR, participants could optionally submit predictions for two auxiliary tasks: language identification (LID) and dialect identification (DID). The LID task involved identifying the spoken language in each utterance, while DID focused on determining the dialect variant. These predictions could be based on the input speech signal or its ASR transcript. Submissions for both auxiliary tasks were accepted for each leaderboard, encouraging the development of end-to-end and multitask ASR systems capable of dialect-aware processing.

For more details on dataset partitions, baseline systems, and evaluation scripts, participants were directed to the challenge website.[1]

III. Datasets

The MADASR 2.0 Challenge is based on the RESPIN corpus, a large-scale multilingual and multidialectal speech dataset covering eight Indian languages and 33 dialects. The RESPIN corpus itself consists of read-speech recordings collected in diverse real-world acoustic conditions. For this challenge, an additional spontaneous speech test set is provided to evaluate system robustness under more natural conversational settings. Each utterance is annotated with a language identifier (LID), a dialect identifier (DID), and corresponding metadata including speaker ID, transcription, and duration. The transcriptions strictly avoid non-native characters, represent all numerical values in textual form, and transliterate English words into the corresponding native scripts. Beyond these constraints, no normalization or standardization is applied to the text, thereby preserving dialectal variations in pronunciation, vocabulary, and orthographic usage.

A. Languages and Dialects

The RESPIN corpus provides fine-grained dialectal labels by associating each utterance with a district-level dialect tag. These dialects span across the eight target languages: Bengali (bn), Bhojpuri (bh), Chhattisgarhi (ch), Kannada (kn), Magahi (mg), Maithili (mt), Marathi (mr), and Telugu (te).

B. Data Composition and Statistics

The subset of the RESPIN dataset used in this challenge is divided into five subsets: *train-small*, *train-large*, *dev*, *test-read*, and *test-spontaneous*. The training subsets contain read speech only, while evaluation is performed on both read and spontaneous test sets. This design enables robust benchmarking under both matched and mismatched conditions. The *train-small* subset (approximately 30 h per language) is used in Tracks 1 and 3, while *train-large* (approximately 150 h per language) is used in Tracks 2 and 4.

Table I presents detailed statistics for each split and language. Metrics include total duration (in hours), number of utterances, number of unique sentences, and speaker counts. Overall, the MADASR 2.0 challenge dataset comprises over 1200 hours of annotated speech across 8 languages, with more than 9000 speakers represented in the training sets. Evaluation sets maintain balanced speaker and dialect coverage across both read and spontaneous conditions.

Additional statistics such as domain-wise distribution, dialectal coverage per split, speaker gender distribution, and vocabulary composition are included in the metadata shared alongside the dataset.[2] These metadata enable deeper analysis of model performance across linguistic, dialectal, and acoustic dimensions.

IV. Challenge Setup and Submission

A dedicated React-based web portal was developed for result submissions. Each team registered with a unique *team name* and was assigned a secure password. Using these credentials, participants could upload their system outputs for any of the four challenge tracks.

[1] https://sites.google.com/view/respinasrchallenge2025/home

[2] https://ee.iisc.ac.in/madasrdataset/

TABLE I
CORPUS STATISTICS ACROSS SUBSETS AND SPLITS FOR ALL EIGHT LANGUAGES IN THE RESPIN DATASET.

		Train-Small				**Train-Large**				**Dev**				**Test-Read**				**Test-Spon**			
LID	#Dials	Dur	#Utts	#Sents	#Spks	Dur	#Utts	#Sents	#Spks	Dur	#Utts	#Sents	#Spks	Dur	#Utts	#Sents	#Spks	Dur	#Utts	#Sents	#Spks
bh	3	27.71	19056	19056	924	142.98	95280	19056	1079	2.14	1500	575	60	3.10	2220	694	120	0.75	655	655	655
bn	5	27.76	17160	17160	997	142.96	85800	17160	1184	2.27	1500	494	100	3.26	2174	648	200	0.75	573	573	573
ch	4	33.82	17160	17160	1041	175.22	85800	17160	1237	2.40	1413	511	80	3.85	2234	695	160	0.99	1147	1147	1147
kn	5	32.46	17160	17160	1331	164.83	85800	17160	1563	2.37	1430	518	100	3.61	2161	663	200	1.00	813	813	813
mg	4	30.72	19056	19056	1162	157.77	95280	19056	1357	2.10	1431	494	80	3.17	2193	640	160	0.96	739	739	739
mr	4	27.46	19056	19056	1377	140.49	95280	19056	1742	1.98	1386	509	80	3.04	2170	711	160	0.88	458	458	458
mt	4	30.60	19056	19056	1374	159.32	95280	19056	1745	2.06	1409	693	80	3.33	2172	993	160	0.75	501	501	501
te	4	30.54	19056	19056	1188	155.89	95280	19056	1415	2.30	1438	500	80	3.37	2226	652	160	0.90	514	514	514
Total	33	241.07	146760	146760	9387	1239.45	733800	146760	11315	17.62	11507	4294	660	26.74	17550	5696	1320	6.97	5400	5400	5400

LID: Language ID; **Dur**: duration in hours; **#Utts**: number of utterances; **#Sents**: number of unique sentences; **#Spks**: number of speakers

TABLE II
ACCEPTED SUBMISSION FORMATS WITH BHOJPURI DECODING EXAMPLES.

Case	**Format and Example**
ASR Only	`281474977512428` बायोगैस टन से बनत बाटे
ASR + LID	`281474977512428` [bh] बायोगैस टन से बनत बाटे
ASR + LID + DID	`281474977512428` [bh_D1] बायोगैस टन से बनत बाटे

Submissions were accepted in tab-separated value (`.tsv`) format and evaluated automatically via a backend API that computed relevant metrics and updated the public leaderboard.

A. Web Portal and Submission Format

Depending on the inclusion of auxiliary task predictions, participants were required to follow one of the three submission formats shown in Table II.

Participants could optionally containerize their evaluation setup using Docker or APIs to support hidden test set evaluation, though this was not mandatory.

B. Challenge Timeline

Table III summarises the official schedule for the MADASR 2.0 Challenge. All deadlines follow the "anywhere on Earth" (AoE) convention. Further details and submission instructions are available on the official challenge website [25].

TABLE III
KEY DATES FOR THE MADASR 2.0 CHALLENGE.

Event	**Date (AoE)**
Registration Opens	April 10, 2025
Training + Dev Set Release	April 19, 2025
Baseline Systems Release	April 22, 2025
Test Set Release	May 25, 2025
Submission Portal Opens	May 31, 2025
Final Submission Deadline	June 21, 2025
Leaderboard Results Announced	June 22, 2025
Challenge Paper Deadline	June 25, 2025

C. Evaluation Metrics and Scoring

ASR performance was measured using character error rate (CER) and word error rate (WER), computed with the `jiwer` toolkit. A small set of invalid utterances was excluded, and any leading LID/DID tokens were stripped before scoring. As the test transcripts contained no punctuation except dots in acronyms, participants were instructed to remove punctuation from their hypotheses; beyond this, no normalization or modification was applied to ensure fairness. Results were reported both overall and language-wise.

LID and DID were evaluated at the utterance level using accuracy. For LID, two-letter language tokens (e.g., `[bn]`) were required; for DID, full dialect tokens (e.g., `[bn_D3]`). Predictions were counted correct only if they exactly matched the reference. DID scoring was restricted to submissions in the "ASR+LID+DID" format.

Three submission formats were supported: (i) `ASR Only`, (ii) `ASR+LID`, and (iii) `ASR+LID+DID`. The evaluation script automatically detected the format via regular expressions, rejected mixed or malformed outputs, and generated: (i) language-wise and overall CER/WER, (ii) LID/DID accuracy (if applicable), and (iii) number of evaluated utterances. All results were automatically logged to the public leaderboard.

This protocol ensured fair, standardised evaluation of transcription accuracy and auxiliary classification, while enabling detailed analysis across languages and dialects.

V. BASELINE SYSTEMS AND OBSERVATIONS

We release four open-source baseline systems,[3] developed using a Conformer encoder and Transformer decoder [26] within a hybrid CTC/attention framework implemented in ESPnet [27]. Tracks 1 and 2 involve supervised training on the 30-hour and 150-hour RESPIN subsets, respectively. Tracks 3 and 4 use the same subsets as Tracks 1 and 2, but initialise the encoder with frozen self-supervised features from IndicWav2Vec.[4] All baseline recipes are publicly available on GitHub, and pretrained model checkpoints for each track are hosted on Hugging Face[5], enabling reproducibility and further research on dialect-aware ASR in Indian languages.

Table IV presents the overall character error rate (CER), word error rate (WER), and language identification (LID) ac-

[3]GitHub repository: https://github.com/saurabhk0317/espnet_respin_asru25/tree/respin_asru25/egs2/respin_asru25/asr1

[4]https://github.com/AI4Bharat/IndicWav2Vec

[5]Track 1–4 models: track1,track2, track3, track4

curacy on the development set. Table V shows the breakdown per language.

TABLE IV
BASELINE RESULTS ON THE DEVELOPMENT SET ACROSS ALL TRACKS.

Track	CER	WER	LID Accuracy
Track 1	4.06	17.28	97.41
Track 2	3.61	15.72	96.58
Track 3	4.36	18.45	96.92
Track 4	3.89	16.74	96.18

As shown in Table IV, **Track 2** achieves the lowest overall CER (3.61%) and WER (15.72%) among all tracks, highlighting the benefit of increased in-domain supervision using 150 hours of labelled read speech. This confirms the importance of corpus scale for effective ASR in dialect-rich scenarios.

Track 4, which combines frozen SSL representations with the same 150-hour dataset, achieves slightly higher CER (3.89%) and WER (16.74%) compared to Track 2. While still competitive, this suggests that in high-resource settings, frozen SSL features alone may not provide additional gains without further adaptation.

Track 3, which pairs frozen SSL features with limited (30-hour) supervision, performs worse than Track 1 in several languages (e.g., `bn`, `kn`, `mt`), reinforcing a known limitation i.e., in low-resource conditions, frozen representations without fine-tuning may not generalise effectively to dialectal and domain-specific variation.

TABLE V
LANGUAGE-WISE BASELINE CER AND WER ON THE DEVELOPMENT SET FOR ALL TRACKS.

LID	Track 1		Track 2		Track 3		Track 4	
	CER	WER	CER	WER	CER	WER	CER	WER
bh	3.86	14.84	3.59	13.77	4.12	15.61	3.94	15.01
bn	4.60	18.35	4.11	16.38	5.01	20.11	4.49	17.70
ch	3.00	11.48	2.61	10.06	3.36	12.43	2.81	10.71
kn	4.44	23.58	3.94	21.95	4.71	24.57	4.27	22.89
mg	5.29	19.18	4.59	17.26	5.55	20.64	4.88	18.24
mr	3.36	16.10	2.76	13.50	3.63	17.21	2.99	14.59
mt	4.43	16.99	4.36	17.34	4.79	18.48	4.64	18.05
te	3.88	22.13	3.41	19.64	4.25	23.72	3.79	21.46

Language-wise results show consistent improvements with larger training sets, while Kannada and Telugu exhibit relatively higher WERs, likely due to script complexity or intra-language variation. These findings affirm the effectiveness of Conformer-based models and the complementary role of self-supervised pretraining and supervised fine-tuning in multilingual, multidialectal ASR.

VI. Overview of Submitted Systems

The MADASR 2.0 challenge attracted substantial attention, with 171 dataset downloads from 28 countries, many beyond those who formally registered or expressed interest. A total of 80 teams from 8 countries officially registered to participate.

Figure 1 shows the geographic distribution of dataset downloads, with India accounting for the majority (55%), followed by Andorra (8.2%), China (5.8%), Bangladesh (6.4%), and the United States (4.7%).

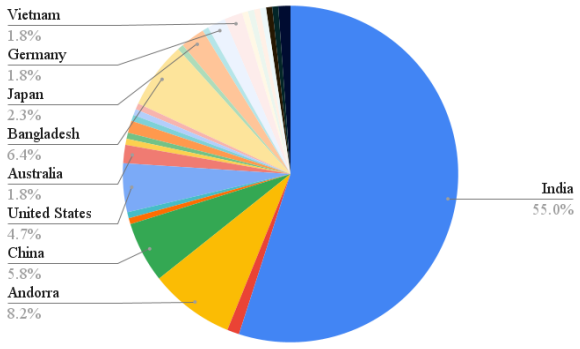

Fig. 1. Geographic distribution of dataset downloads across 28 countries

Ultimately, 7 unique teams from 2 countries submitted at least one valid result across one or more tracks. The challenge received a total of 306 submissions, out of which 191 met the format and evaluation requirements and were accepted for leaderboard ranking. Track-wise participation varied, with the highest activity observed in Tracks 1 and 3. Table VI summarises these statistics.

The list of participating organisations and team names used in official results is available on the challenge website.[6]

TABLE VI
PARTICIPATION AND SUBMISSION STATISTICS FOR MADASR 2.0. NUMBERS IN PARENTHESES INDICATE THE NUMBER OF COUNTRIES.

Item	Count
# Data downloads	171 (28)
# Interested participants	89 (10)
# Registered teams	80 (8)
# Track 1:2:3:4 registrants	62:44:55:51
# Unique final submissions	7
# Submissions per track (1:2:3:4)	138:29:132:7
# Valid submissions per track (1:2:3:4)	78:28:78:7
Total submissions	306
Total valid submissions	191

A. Representative Approaches

We summarise key approaches from four representative teams that submitted high-performing systems to the MADASR 2.0 Challenge.

a) Team YS: This team fine-tuned Whisper-large-v3 [2] using *MixLoRA*, a parameter-efficient method combining low-rank adaptation with top-k expert routing [28]. MixLoRA reduced trainable parameters while preserving capacity, enabling efficient adaptation to the linguistic diversity in RESPIN.

[6]https://sites.google.com/view/respinasrchallenge2025/home

The approach outperformed conventional fine-tuning on read speech while maintaining computational efficiency.

b) Team SPRING_Lab: To address script overlap and phoneme similarity across Indian languages, the team used a grapheme-to-phoneme (G2P) mapping to convert all transcripts into a *common label set* (CLS) [29], [30]. A dual-decoder model was trained: one decoder predicted CLS outputs, and another translated them back to the native script. The second decoder, with cross-attention to the encoder, implicitly learned language and dialect cues. This method improved WER, LID, and DID accuracy, with additional comparisons against a cascaded CLS-to-script model and BPE-based tokenisations.

c) Team QWER: Under Track 1's constrained setup, this team explored four strategies: (1) voice conversion for data augmentation, (2) mixed-speech augmentation with low-volume background speech, (3) insertion of LID tokens at transcript boundaries, and (4) error-aware fine-tuning by upsampling high-error dialects. Based on dev performance, the latter three were used for final evaluation, leading to robust ASR results in the low-resource setting.

d) Team Pramiteeh: This team developed a unified multitask model for ASR, LID, DID, gender, and age-group prediction. A shared encoder fed into task-specific heads, with language-specific ASR decoders using distinct vocabularies. A dynamic routing mechanism directed utterances to the appropriate decoder, enabling efficient multilingual modelling. The model was trained end-to-end with a composite loss and fine-tuned on RESPIN after pretraining with speed-perturbed audio, yielding strong performance across all tasks.

VII. Leaderboard Results and Analysis

Table VII presents the final system leaderboard for the MADASR 2.0 Challenge, summarising CER, WER, LID/DID accuracy, and language-wise CER across tracks and test sets. We discuss trends and key observations across read and spontaneous speech conditions, language-specific performance, and auxiliary tasks.

A. Read Speech Test Set

The read speech test set saw strong ASR performance, with CERs ranging from 4.30% to 8.07%. In Track 1, QWER (4.84%) and the baseline (4.86%) performed comparably, with the baseline achieving the highest LID accuracy (97.08%). SPRING_Lab reported a slightly higher CER (5.57%) but demonstrated strong LID (96.80%) and the best DID accuracy (70.80%). Pramiteeh had a higher CER (6.66%) but maintained competitive LID performance (96.09%).

Track 2 showed the overall best CERs. The baseline (4.30%) and SPRING_Lab (4.40%) were closely matched, with SPRING_Lab achieving the highest LID (97.39%) and DID (75.36%) scores. Pramiteeh followed with 6.07% CER and strong LID accuracy (97.17%).

In Track 3, YS (4.98%) and the baseline (5.36%) performed well. Pramiteeh (6.60%) showed good LID (96.89%) and DID (67.37%) performance. Whissle (17.33%) and AR_India (50.89%) underperformed significantly, indicating potential issues in their use of external resources.

Track 4 results showed the baseline leading with 4.68% CER and 95.15% LID accuracy, followed by Pramiteeh (6.71%, 96.00%) and Whissle (8.07%, 96.44%). These results suggest that when external resources are combined with sufficient in-domain data, systems can maintain robust performance.

B. Spontaneous Speech Test Set

Performance dropped notably on the spontaneous test set due to increased acoustic and linguistic variability. CERs ranged from 23.66% to 67.52%, with wider gaps across systems compared to the read speech setup.

In Track 1, Pramiteeh (23.67%) achieved the best CER, followed closely by the baseline (25.39%). QWER (26.16%) showed significantly lower LID accuracy (21.20%), while SPRING_Lab had the highest LID (72.68%) and the only DID submission (29.08%).

Track 2 results were more stable, with Pramiteeh (24.50%) again outperforming the baseline (25.11%) and SPRING_Lab (26.99%). SPRING_Lab also reported the highest LID (77.61%) and DID (33.33%) scores in this track.

In Track 3, Pramiteeh matched its earlier CER (23.66%), while the baseline and YS trailed slightly (25.99%, 26.51%). Whissle (40.00%) and AR_India (67.52%) showed substantial degradation, reflecting the difficulty of generalising external models to spontaneous speech.

Track 4 showed moderate variation: Pramiteeh (24.50%) and the baseline (24.94%) had comparable performance, while Whissle (28.48%) showed slightly higher error. LID accuracies remained high across all systems ($\geq$76.76%), with Pramiteeh reporting DID accuracy of 26.52%.

These results reinforce the importance of in-domain training for spontaneous speech and highlight the challenges of domain mismatch even with auxiliary supervision.

C. Language-wise and Auxiliary Task Performance

Language-wise CER patterns reveal consistent trends across tracks. Languages such as Kannada (kn), Magahi (mg), and Telugu (te) reported higher CERs, especially in spontaneous speech, indicating greater modelling difficulty. In contrast, Bhojpuri (bh), Bengali (bn), and Maithili (mt) consistently achieved lower CERs, suggesting easier recognition characteristics under current models.

Among teams that submitted auxiliary task predictions, Pramiteeh consistently achieved high LID accuracy ($\geq$96%) across all tracks and strong DID accuracy, particularly in Track 2 (71.83%). SPRING_Lab's dual-decoder and CLS-token approach yielded robust DID performance, peaking at 75.36% in Track 2.

D. Key Takeaways

- **Track 2** (constrained high-resource) achieved the best overall ASR performance, highlighting the value of increased in-domain training data.

TABLE VII
SYSTEM LEADERBOARD ACROSS TRACKS AND TEST SETS, SHOWING OVERALL CER/WER, LANGUAGE-WISE CER, AND LID/DID ACCURACIES WHERE APPLICABLE.

Team	Overall Error-rates (%)		Classification Accuracies (%)		Language-wise CER (%)							
	CER	WER	LID	DID	bh	bn	ch	kn	mg	mr	mt	te
Read Speech Test Set												
Track 1												
QWER	4.84	18.68	12.73	NA	4.49	4.79	3.62	5.29	6.32	3.96	5.41	4.92
baseline	4.86	18.70	97.08	NA	4.56	4.80	3.52	5.45	6.44	4.05	5.27	4.91
SPRING_Lab	5.57	21.09	96.80	70.80	4.86	5.92	4.25	6.13	7.01	4.64	6.04	5.76
pramiteeh	6.66	24.43	96.09	NA	6.18	9.80	5.32	6.56	8.03	4.56	7.24	5.78
Track 2												
baseline	4.30	16.92	96.03	NA	4.14	4.19	3.15	4.70	5.63	3.28	5.04	4.40
SPRING_Lab	4.40	17.06	97.39	75.36	4.11	4.60	3.24	4.93	5.57	3.39	4.93	4.47
pramiteeh	6.07	22.32	97.17	NA	5.85	8.91	4.83	5.84	7.29	4.17	6.62	5.23
Track 3												
YS	4.98	18.83	96.54	NA	4.36	5.23	3.47	5.76	6.40	3.78	5.55	5.32
baseline	5.36	20.32	96.63	NA	4.87	5.38	3.97	5.85	7.06	4.40	5.84	5.57
pramiteeh	6.60	24.02	96.89	67.37	5.96	9.25	6.01	6.47	7.76	4.67	6.82	5.97
whissle	17.33	52.15	86.28	NA	11.64	26.28	11.58	20.20	14.58	18.61	13.95	20.63
AR_India	50.89	77.81	NA	NA	48.64	52.61	49.29	51.23	50.46	51.53	51.33	51.85
Track 4												
baseline	4.68	18.14	95.15	NA	4.27	4.75	3.55	5.11	5.99	3.53	5.46	4.86
pramiteeh	6.71	24.53	96.00	62.58	6.33	9.45	5.42	6.53	8.10	5.12	7.26	5.72
whissle	8.07	29.67	96.44	NA	7.83	12.37	5.97	8.19	8.41	6.61	7.82	7.44
Spontaneous Speech Test Set												
Track 1												
pramiteeh	23.67	59.80	80.02	NA	24.85	24.29	20.36	27.69	27.51	13.06	27.10	22.02
baseline	25.39	61.70	79.94	NA	26.08	25.64	20.52	30.74	27.32	16.06	27.33	26.28
QWER	26.16	61.67	21.20	NA	27.78	28.11	20.74	31.33	28.34	15.71	27.80	26.65
SPRING_Lab	30.63	67.01	72.68	29.08	29.63	39.03	23.29	40.39	30.27	17.82	30.10	29.70
Track 2												
pramiteeh	24.50	61.00	77.61	26.52	25.35	25.84	20.74	28.90	27.85	14.29	27.34	23.16
baseline	25.11	59.01	76.89	NA	25.50	30.06	19.97	30.86	26.10	14.37	25.59	25.34
SPRING_Lab	26.99	62.32	77.61	33.33	27.08	31.39	20.98	33.10	27.55	15.26	28.68	28.63
Track 3												
pramiteeh	23.66	59.79	80.02	NA	24.85	24.29	20.35	27.69	27.51	13.06	27.10	22.00
baseline	25.99	62.73	77.50	NA	28.22	26.35	20.75	32.15	29.16	14.06	27.68	25.82
YS	26.51	58.55	75.59	NA	25.79	29.69	19.13	34.50	27.64	12.55	26.01	31.51
whissle	40.00	79.36	70.88	NA	37.24	46.66	33.25	50.26	37.57	30.56	35.37	43.10
AR_India	67.52	95.33	NA	NA	73.20	64.50	63.88	67.24	69.69	61.82	71.17	69.15
Track 4												
pramiteeh	24.50	61.00	77.61	26.52	25.35	25.84	20.74	28.90	27.85	14.29	27.34	23.16
baseline	24.94	59.59	76.00	NA	26.09	26.57	19.73	31.58	27.14	13.30	26.43	24.80
whissle	28.48	66.09	76.76	NA	31.31	33.76	21.47	34.14	29.66	16.20	28.93	29.62

- **Spontaneous speech** led to substantial CER/WER degradation across all systems, revealing the challenge of domain and style mismatch.
- **External resources** (Tracks 3 and 4) showed mixed impact, only some systems benefited, while others were hindered by poor domain adaptation.
- **Multi-task learning** (Pramiteeh) and **script-agnostic modeling** (SPRING_Lab) enhanced auxiliary task performance without hurting ASR accuracy.

VIII. CONCLUSION

MADASR 2.0 advances research in multilingual and multidialectal ASR by releasing a large-scale benchmark derived from the RESPIN corpus, featuring 8 languages, 33 dialects, and both read and spontaneous speech. Organised across four tracks with varying resource constraints and optional LID/DID tasks, the challenge attracted diverse modelling approaches. Results show that ASR on read speech is approaching maturity, while spontaneous speech remains challenging due to domain mismatch and acoustic variability. Multitask learning, parameter-efficient adaptation, and data augmentation emerged as effective strategies. MADASR 2.0 provides a standardised platform for evaluating dialect-aware models and lays the groundwork for future research in inclusive ASR for linguistically diverse populations.

ACKNOWLEDGMENT

This work was supported by the RESPIN project, funded by the Gates Foundation. We thank the RESPIN team and our project partner, Navana Tech, for their valuable support in data collection and preparation.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "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, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[3] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Interspeech 2020*, 2020, pp. 4751–4755.

[4] T. Javed, S. Doddapaneni, A. Raman, K. S. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Towards building asr systems for the next billion users," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 813–10 821.

[5] H. Palivela, M. Narvekar, D. Asirvatham, S. Bhushan, V. Rishiwal, and U. Agarwal, "Code-Switching ASR for Low-Resource Indic Languages: A Hindi-Marathi Case Study," *IEEE Access*, vol. 13, pp. 9171–9198, 2025.

[6] T. Khan and M. K. Singh, "Variations in Bhojpuri: A sociolinguistic study," *Linguistic Ecology of Bihar*, 2019.

[7] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, and K. Sankaranarayanan, "MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," in *Proc. Interspeech*, 2021, pp. 2446–2450.

[8] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 5123671, 2021.

[9] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on deep learning for speech recognition and related applications*. Citeseer, 2009, pp. 1–4.

[10] S. Punjabi, H. Arsikere, Z. Raeesy, C. Chandak, N. Bhave, A. Bansal, M. Müller, S. Murillo, A. Rastrow, A. Stolcke *et al.*, "Joint ASR and language identification using RNN-T: An efficient approach to dynamic language switching," in *Proc. ICASSP*. IEEE, 2021, pp. 7218–7222.

[11] A. Lyu, Z. Wang, and H. Zhu, "Ant Multilingual Recognition System for OLR 2021 Challenge," in *Proc. Interspeech*, 2022, pp. 3684–3688.

[12] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, and S. Watanabe, "Improving Massively Multilingual ASR with Auxiliary CTC Objectives," in *Proc. ICASSP*, 2023, pp. 1–5.

[13] Shaik, Mohammed Maqsood and Klakow, Dietrich and Abdullah, Badr M., "Self-Supervised Adaptive Pre-Training of Multilingual Speech Models for Language and Dialect Identification," in *Proc. ICASSP*, 2024, pp. 11 436–11 440.

[14] M. C. S. Priya, D. K. Renuka, L. A. Kumar, and S. L. Rose, "Multilingual low resource Indian language speech recognition and spell correction using Indic BERT," *Sādhanā*, vol. 47, no. 227, 2022.

[15] A. Arunkumar, M. D. Batra, and S. Umesh, "DuDe: Dual-Decoder Multilingual ASR for Indian Languages using Common Label Set," *ArXiv*, vol. abs/2210.16739, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253237346

[16] C. Zhang, B. Li, T. Sainath, T. Strohman, S. Mavandadi, S.-Y. Chang, and P. Haghani, "Streaming End-to-End Multilingual Speech Recognition with Joint Language Identification," in *Proc. Interspeech*, 2022, pp. 3223–3227.

[17] L. Zhou, J. Li, E. Sun, and S. Liu, "A Configurable Multilingual Model is All You Need to Recognize All Languages," in *Proc. ICASSP*, 2022, pp. 6422–6426.

[18] Amartyaveer, S. Kumar, S. Sharma, S. Udupa, S. Badiger, A. Singh, D. G, J. Bandekar, S. Murthy, and P. Kumar Ghosh, "Improving Dialect Identification in Indian Languages Using Multimodal Features from Dialect Informed ASR," in *Proc. ICASSP*, 2025, pp. 1–5.

[19] S. Kumar, Amartyaveer, and P. K. Ghosh, "Jointly Improving Dialect Identification and ASR in Indian Languages using Multimodal Feature Fusion," in *Accepted at Interspeech*, 2025.

[20] MADASR Challenge Organizers, "MADASR: Model Adaptation for ASR in Low-resource Indian Languages," https://sites.google.com/view/respinasrchallenge2023, 2023, accessed: June 23, 2025.

[21] A. Singh, A. S. Mehta, J. Nanavati, J. Bandekar, K. Basumatary, S. Badiger, S. Udupa, S. Kumar, P. K. Ghosh, P. Pai *et al.*, "Model Adaptation for ASR in low-resource Indian Languages," *arXiv preprint arXiv:2307.07948*, 2023.

[22] S. Udupa, J. Bandekar, G. Deekshitha, S. Kumar, P. K. Ghosh, S. Badiger, A. Singh, S. Murthy, P. Pai, S. Raghavan, and R. Nanavati, "Gated Multi Encoders and Multitask Objectives for Dialectal Speech Recognition in Indian Languages," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[23] T. Alumäe, J. Kong, and D. Robnikov, "Dialect Adaptation and Data Augmentation for Low-Resource ASR: Taltech Systems for the Madasr 2023 Challenge," in *Proc. IEEE ASRU*, 2023, pp. 1–7.

[24] RESPIN Team, "RESPIN: Speech Recognition in Agriculture and Finance for the Poor in India," https://respin.iisc.ac.in/, 2024, accessed: September 2, 2024.

[25] MADASR 2.0 Challenge Organizers, "MADASR 2.0: Multi-Lingual Multi-Dialect ASR in 8 Indian Languages," https://sites.google.com/view/respinasrchallenge2025, 2025, accessed: June 23, 2025.

[26] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[27] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[28] D. Li, Y. Ma, N. Wang, Z. Ye, Z. Cheng, Y. Tang, Y. Zhang, L. Duan, J. Zuo, C. Yang, and M. Tang, "MixLoRA: Enhancing Large Language Models Fine-Tuning with LoRA-based Mixture of Experts," 2024. [Online]. Available: https://arxiv.org/abs/2404.15159

[29] A. Prakash, A. Leela Thomas, S. Umesh, and H. A Murthy, "Building Multilingual End-to-End Speech Synthesisers for Indian Languages," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 194–199.

[30] V. M. Shetty and S. Umesh, "Exploring the use of Common Label Set to Improve Speech Recognition of Low Resource Indian Languages," in *Proc. ICASSP*, 2021, pp. 7228–7232.