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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce **RESPIN-S1.0**, the largest publicly available dialect-rich read speech corpus for Indian languages, comprising over 10,000 hours of validated audio spanning nine major languages – Bengali, Bhojpuri, Chhattisgarhi, Hindi, Kannada, Magahi, Maithili, Marathi, and Telugu. Indian languages are characterized by high dialectal variation and are spoken by populations that are often digitally underserved. Existing speech corpora typically represent only standard dialects and lack domain relevance. RESPIN-S1.0 fills this critical gap by collecting speech across 38+ dialects and two high-impact domains: agriculture and finance. Text data was carefully composed by native dialect speakers and validated via a robust pipeline involving both automatic and manual checks. Over 200,000 utterances were recorded through a crowdsourced mobile application by native speakers and subsequently categorized into clean, semi-noisy, and noisy slabs based on transcription quality. The clean slab alone exceeds 10,000 hours. RESPIN also provides speaker metadata, phonetic lexicons, and dialect-aware train-dev-test splits to ensure reproducibility. To benchmark performance, we evaluate a range of ASR models – TDNN-HMM, E-Branchformer, Whisper, IndicWav2Vec2, and SPRING SSL models – and find that fine-tuning on RESPIN significantly improves recognition accuracy over existing pretrained models. A subset of RESPIN-S1.0 has already supported community efforts through challenges such as the SLT Code Hackathon 2022 and MADASR@ASRU 2023/2025, with over 1200 hours of data released publicly. This resource supports research in dialectal ASR, LID, DID, and speech-related areas, and sets a new standard for inclusive, dialect-rich corpora in multilingual, low-resource settings.

1 Introduction

India's vast linguistic diversity – with 22 scheduled languages and hundreds of dialects [1] – necessitates speech technologies that support local languages to ensure inclusivity. However, development in this space remains limited due to the scarcity of curated audio-text datasets [1], especially for dialectal variation [2]. Roughly 64% of Indias population lives in rural areas, and 57.8% belong to agricultural households [3], yet ASR research has largely focused on English or standard language forms [4]. Existing corpora often cover only standard dialects [2, 5], leading to poor performance on regionally diverse speech.

To address this, RESPIN-S1.0 introduces a large-scale, multi-dialectal, multi-domain read speech corpus for nine Indian languages – Bengali, Bhojpuri, Chhattisgarhi, Hindi, Kannada, Magahi,

[1] `https://censusindia.gov.in/nada/index.php/catalog/42561/download/46187/Language_Atlas_2011.pdf`

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

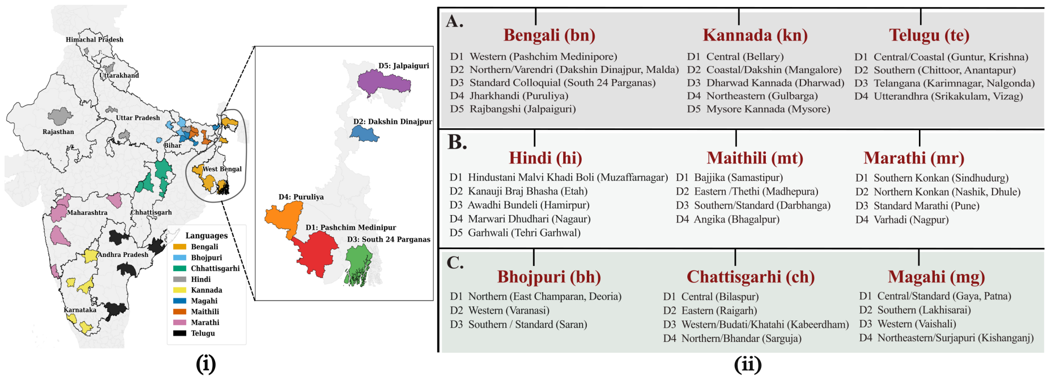

Figure 1: (i) District-level distribution of the nine RESPIN languages across India, based on the 2011 Census (illustrative, not to scale). Each language is shown in a distinct color. A sample inset shows dialect-wise representation for Bengali. (ii) Language classification: A – Scheduled, non-Devanagari; B – Scheduled, Devanagari; C – Non-scheduled, Devanagari.

Maithili, Marathi, and Telugu. Languages were selected based on speaker population, socio-economic indicators, and availability of resources. Figure 1 shows the district-level language distribution and dialect breakdown. RESPIN is the first public corpus to provide large-scale dialectal data for Bhojpuri, Chhattisgarhi, and Magahi. The pipeline – from sentence composition to audio validation – was conducted at the dialect level to preserve linguistic integrity. It also includes manually verified phonetic lexicons (as per ILSL guidelines [6]) and rich speaker metadata (e.g., pincode, gender, age group).

To promote reproducibility, RESPIN provides train/dev/test splits, dialect-level metadata, and ASR benchmarks using TDNN-HMM [7], E-Branchformer [8], Whisper [9], and SSL models like IndicWav2Vec2 [10] and SPRING-Data2Vec [2]. Fine-tuning on RESPIN consistently improves ASR performance over models trained on external corpora. RESPIN has already enabled multilingual ASR research through the SLT Code Hackathon 2022 [3] and MADASR challenges (ASRU 2023, ASRU 2025) [4], where 1200+ hours were released to the participants. By capturing Indias linguistic depth, RESPIN advances inclusive voice technologies for underserved communities in the Global South.

2 Background

Table 1: Existing Indic Datasets

Dataset	Languages	Domains	Districts	Hours	Speakers	Sentences	Source
INDICVOICES [11]	13	52	145	7348	16237	11,00,000+	Wikipedia, Composed, Spontaneous
INDICVOICES-R [12]	22	multi	multi	1704	10496	NA	NA
Kathbath [13]	12	multi	203	1684	1217	12,00,000+	IndicCorp (Web data)
Shrutilipi [14]	12	multi	NA	6457	NA	33,00,000	All India Radio
NPTEL [15]	8	1	NA	857	NA	NA	Lectures
Svarah [16]	1	9	65	9.6	117	NA	Wikipedia, Prompts, Spontaneous
SPRING-INX [17]	10	multi	40+	2000	7609	NA	NA
SPIRE-SIES [18]	1	NA	NA	193	1607	NA	NA
FLEURS [19]	13	NA	NA	156	39	NA	Wikipedia
Gram Vaani [20]	1	multi	25	1108	NA	NA	Spontaneous Speech
IISc-MILE [21]	2	NA	NA	497	1446	NA	NA
MUCS [21]	3	4	4 (for Odia)	NA	310	9080	NA
Vāksāncayah [22]	1	8	NA	78	27	46,000	Online stories
E&NE languages [23]	4	NA	multi	19.75	NA	NA	NA
NISP [24]	6	NA	NA	56.86	345	NA	news, TIMIT
CommonVoice [25]	8?	4?	NA	373?	NA	NA	Wikipedia, Composed
CMS [26]	6	NA	NA	35	243	NA	Composed
IITB-MSC [27]	1	1	1	109	36	3000	Textbooks
IndicSpeech [28]	3	NA	NA	24	3	42,046	Online news
MSR Challenge [29]	3	NA	NA	150	1286	1,02,397	NA
Google TTS [30]	1	NA	NA	3	6	NA	NA
IIITH-ILSC [31]	23	NA	NA	103.5	1150	NA	NA
IndicTTS [32]	13	4+	NA	389.6	26	NA	Literature, newspapers
IIITH-ISD [33]	7	NA	NA	11	35	1000	Wikipedia
RESPIN-S1.0	**9**	**2**	**38+**	**10,416.58**	**18,000+**	**2,09,822**	**Composed**

NA = Information Not Available

[2] https://asr.iitm.ac.in/models

[3] https://sites.google.com/view/slt-team

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

Table 1 compares major open-source Indic speech corpora across languages, domains, districts, duration, speaker count, and data sources. While many datasets offer broad language coverage and large volumes of audio, they often lack dialectal diversity and regional specificity – critical for building inclusive ASR systems for rural populations. Most rely on publicly available web content (e.g., Wikipedia, books, news), resulting in generic domain coverage and limited alignment with real-world use. RESPIN-S1.0 addresses these limitations by focusing on agriculture and banking – domains essential to Indias low-literacy and rural communities – and by manually composing 2,09,822 sentences to reflect regionally grounded, colloquial speech. Unlike large-scale efforts like INDICVOICES [11] and INDICVOICES-R [12], which cover scheduled languages, RESPIN also includes low-resource, non-scheduled languages such as Bhojpuri, Chhattisgarhi, and Magahi, often mislabeled as Hindi dialects. With over 10,000 hours of validated audio from 18,000+ speakers across 38+ dialect-rich districts, RESPIN offers the most comprehensive dialect-aware resource for speech technology in Indian languages.

3 Data collection and Validation pipeline

RESPIN is the first large-scale Indian speech corpus to preserve dialectal integrity throughout the data creation process. As shown in Figure 2, the pipeline includes language and dialect selection, manual text composition, multi-stage validation, and speaker-level audio collection. Unlike corpora built from scraped or generic online content, RESPIN focuses on agriculture and finance, with all text and audio created and validated at the dialect level.

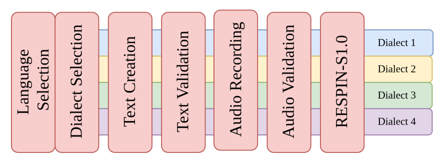

Figure 2: Data creation pipeline maintaining dialectal integrity.

3.1 Language and Dialect Selection

According to the Census of India 2011, 50.58M, 16.25M, 12.71M, and 13.58M people speak Bhojpuri, Chhattisgarhi, Magahi, and Maithili, respectively. While Magahi is often misclassified as a dialect of Hindi, it is a different branch of the Indo-Aryan subfamily. To support such large speaker populations, it is essential to develop robust language models with rich vocabularies and large-scale sentence corpora for each language. RESPIN aims to build an ecosystem of speech recognition resources tailored to empower Indias working-class population. In 2022-23, 45.76% of India's workforce was engaged in agriculture and allied sectors, while finance and banking continue to play a critical role in daily transactions and access to services. By focusing on these two domains, RESPIN seeks to bridge the gap between under-resourced language communities and accessible, voice-driven technologies.

To support domain-specific sentence creation, a comprehensive set of topics was curated across agriculture and finance. This ensured focused guidance for sentence composers, especially those unfamiliar with the subject matter. The topics were manually compiled from diverse sources such as magazines, websites, academic portals, and Wikipedias outline articles. Wikipedias topic trees and linked articles were particularly useful in structuring the coverage. The final list includes around 1500 topics, each associated with relevant web links for reference. Starting with broad categories – such as crop cultivation or digital banking – the list progressively narrows to subtopics, including sugarcane harvesting techniques, UPI PIN setup, and transaction history checks in mobile apps. This curated topic bank ensured comprehensive and relevant coverage of the target domains.

3.2 Text Data Acquisition and Validation

The creation of a dialect-level text corpus was the foundational step in building RESPIN. Figure 3 outlines the overall pipeline. The process began with onboarding and training dialect experts who helped curate text with high dialectal specificity, ensuring the inclusion of regional nuances and speech variation. As discussed earlier, the corpus was designed to collect dialect-rich sentences from agriculture and finance domains – making RESPIN uniquely domain-specific. Native speakers were hired through a multi-stage selection process to compose these sentences. The raw text was then passed through a validation pipeline combining automatic and manual checks to ensure compliance with linguistic guidelines. Only validated sentences were used in the audio data collection phase.

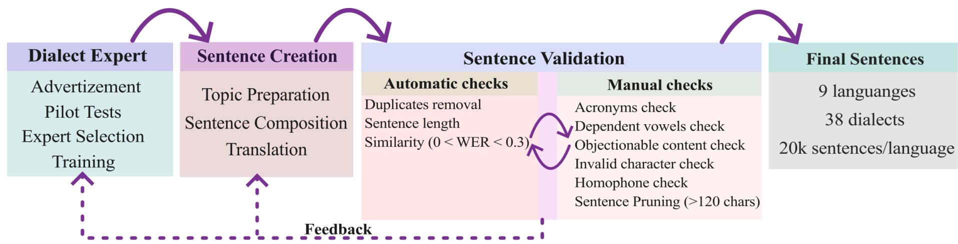

Figure 3: Flowchart showing the RESPIN text data preparation.

3.2.1 Sentence Creation

While large volumes of digital text exist in standard language formats, they often lack colloquial style and dialectal variation. To address this, RESPIN prioritized sourcing sentences directly from native speakers across districts, ensuring that the text reflects authentic regional expressions and resonates with local speech patterns. Composers were tasked with crafting conversational, domain-specific sentences aligned with designated topics in agriculture and finance. This approach enriched the linguistic diversity of the corpus but also posed challenges, as dialectal variation can differ even within a 5 km radius. Recognizing the fluid and non-standardized nature of dialects, we adopted an inclusive strategy that embraced intra-dialectal variation, resulting in a rich and representative dataset.

Sentence composition followed strict guidelines to ensure consistency and usability: limiting sentence length, avoiding sentence-initial pronouns, excluding non-language numerals, restricting punctuation to full-stop (.), comma (,) and question mark (?), avoiding controversial content, adhering to topic relevance, and maintaining standard acronym formatting. Manual composition was the most accurate but also the most resource-intensive data creation method. While it served as the primary strategy, translation from already composed sentences was used to fill gaps in some dialects. The proportion of translated sentences in Bhojpuri, Chhattisgarhi, Hindi, Kannada, Magahi, Maithili, Marathi, and Telugu was 6.65%, 100%, 9.8%, 0.1%, 0.4%, 16.5%, 5.1%, and 5.2%, respectively. Bengali sentences were entirely composed from scratch.

3.2.2 Sentence Validation

The raw composed text corpus is passed through a multi-stage validation pipeline involving both automated (AC) and manual checks (MC) by trained language validators. As multiple contributors are involved in sentence composition, inconsistencies and errors are inevitable. Since the corpus is used as stimuli for crowd-sourced audio recording, each sentence must be accurate, unambiguous, coherent, and compatible with the recording interface, making validation essential. The pipeline architecture is largely consistent across languages but includes language-specific adaptations. Key checks include: (1) duplicate removal (AC), (2) invalid character correction (MC), (3) sentence length pruning (MC), (4) acronym standardization (MC), (5) matra correction (MC), (6) word-level edits (MC), (7) similar sentence filtering (MC), (8) homophone disambiguation (MC), and (9) additional language-specific checks (see Appendix A.1). Approximately 3.6% of the raw corpus was discarded due to unfixable errors or dialect mismatch. The validation process follows a versioned workflow, where each stage produces a new corpus version to enable rollback and auditing. Independent checks are applied within a single version, while dependent checks are performed sequentially.

3.3 Audio Data Acquisition and Validation

Following the validation of dialect-specific text corpus, audio data collection was conducted via a mobile application. Native speakers of each dialect were prompted to read validated sentences aloud and record them in quiet environments. Each speaker was assigned a maximum of 577 sentences. In some cases, speakers recorded additional sentences to meet dialect-wise targets due to dropouts by other participants. To capture intra-dialectal acoustic variation, each sentence was recorded by multiple speakers – typically between 30 and 150. This many-to-one sentence-to-speaker mapping enables the corpus to represent diverse pronunciation styles, prosodic patterns, and speech rates within each dialect.

3.3.1 Audio Validation Pipeline

The recorded audio data underwent a structured validation pipeline combining manual and semi-automated checks. Initially, approximately 5% of the utterances in each dialect were manually audited to assess whether the recorded audio matched the corresponding text. Based on this validation, the entire dataset was partitioned into three quality slabs – *Clean*, *Semi-noisy*, and *Noisy* – using a semi-automated scoring process.

The slab assignment reflects the proportion of audio-text pairs that are exact matches: the clean slab contains the highest percentage of perfectly aligned utterances, while the noisy slab contains the least. This slab-based categorization allows downstream tasks to select data based on quality requirements and robustness needs. Complete definitions of slabs and associated thresholds are included in the supplementary appendix.

This validation framework ensures that the RESPIN audio corpus is high-quality, dialect-specific, and suitable for benchmarking robust ASR systems under realistic multilingual and multi-dialect conditions.

4 RESPIN-S1.0 Corpus

4.1 Text Data Analysis

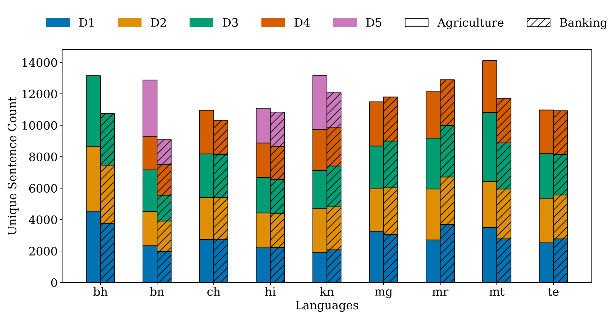

Figure 4: Unique sentence count per dialect, domain and language.

Table 2: Lexicon statistics across languages.

LID	#chars	#phones	#words
bn	64	50	18571
bh	71	54	14105
ch	68	50	13230
hi	72	55	16571
kn	66	50	50822
mg	72	54	21711
mt	72	55	19336
mr	68	51	35709
te	63	48	39235

Figure 4 shows the distribution of unique sentence counts across dialects, domains (Agriculture and Banking), and languages. Each language includes over 20,000 curated sentences across 3-5 dialects, with coverage in both domains per dialect. While perfect balance is constrained by dialect experts' availability and regional factors, the dataset maintains approximate uniformity across dialect-domain pairs. Notable deviations – e.g., higher contributions from dialect D5 in `kn` and D3 in `mt` – likely reflect stronger regional participation or easier contributor access.

Table 2 reports lexicon statistics, including unique characters, phonemes, and words per language. Lexicons are derived from the full sentence set. Kannada (`kn`) and Telugu (`te`) show higher word counts (50k and 39k+), indicative of richer morphology and larger text pools. In contrast, Bhojpuri (`bh`) and Chhattisgarhi (`ch`) have more compact vocabularies, possibly due to lower lexical variation. Character counts (63-72) align with script complexity, and phoneme inventories (50-55) match known Indo-Aryan and Dravidian phonological structures.

Together, these statistics reflect the linguistic richness and dialectal coverage of the text corpus. The balance across dialects and domains, combined with diverse lexicons, makes the dataset a strong foundation for multilingual and multidialectal ASR, language modeling, and speech-language technology research.

Table 3: Dialect-wise duration (in hours) across Clean, Semi-noisy, and Noisy subsets for 9 Indian languages.

Dialect	Type	bh	bn	ch	hi	kn	mg	mr	mt	te
D1	Clean	351.25	206.40	344.89	205.25	237.38	340.88	312.58	195.10	348.78
	Semi-noisy	32.09	64.61	31.75	49.35	58.72	15.12	63.69	117.80	51.61
	Noisy	41.43	1.31	21.07	80.67	52.81	17.60	61.15	71.01	41.96
D2	Clean	417.74	271.45	329.20	159.78	245.03	349.01	328.89	112.16	333.28
	Semi-noisy	11.25	12.97	22.37	90.78	37.07	13.94	54.39	139.60	74.12
	Noisy	5.68	0.80	12.07	88.07	38.36	13.13	49.17	180.44	33.67
D3	Clean	347.97	283.17	297.63	195.93	235.92	333.33	321.62	203.16	331.65
	Semi-noisy	62.53	22.55	77.19	70.51	55.35	26.11	60.99	164.29	58.34
	Noisy	29.46	1.10	22.81	68.28	44.17	14.87	23.49	55.73	54.49
D4	Clean	–	216.14	324.25	138.83	248.10	321.18	316.39	212.64	290.27
	Semi-noisy	–	62.64	67.56	116.41	34.66	57.22	156.14	88.55	38.46
	Noisy	–	2.13	34.17	99.35	48.41	27.14	66.13	98.11	18.87
D5	Clean	–	236.08	–	245.14	228.13	–	–	–	–
	Semi-noisy	–	27.19	–	35.74	64.40	–	–	–	–
	Noisy	–	1.39	–	54.49	42.48	–	–	–	–
Total	Clean	1116.96	1213.24	1295.97	944.93	1194.56	1344.40	1279.48	723.06	1303.98
Total	Semi-noisy	105.87	189.96	198.87	362.79	250.20	112.39	335.21	510.24	222.53
Total	Noisy	76.57	6.73	90.12	390.86	226.23	72.74	199.94	405.29	148.99

4.2 Audio Data Analysis

4.2.1 Slab-Wise Audio Distribution

Table 3 summarizes dialect-wise audio durations (in hours) across the Clean, Semi-noisy, and Noisy slabs for all nine Indian languages. The full corpus contains over 12,000 hours of read-speech audio, covering more than 20,000 sentences per language. Based on transcription quality and alignment confidence (see Section 3.3.1), audio is grouped into three slabs: *Clean*, *Semi-noisy*, and *Noisy*.

The intended target was 200 hours of Clean data per dialect for languages with five dialects (e.g., Hindi, Bengali, Kannada), and 250 hours per dialect for those with four dialects (e.g., Magahi, Marathi, Telugu). While most dialects met this target – particularly in Bengali, Chhattisgarhi, Kannada, and Marathi – some under-resourced dialects (e.g., Hindi D2, D4 and Maithili D2) fell short, requiring larger proportions of *Semi-noisy* and *Noisy* data to ensure sufficient representation. Such shortfalls are likely due to challenges in recruiting fluent readers in specific dialects, influenced by literacy, regional accessibility, and dialectal overlap. For example, Maithili and Hindi have lower *Clean*-slab totals – 723.06 and 944.93 hours respectively – compared to other languages that exceed 1100 hours.

Across the full dataset, the *Clean* slab comprises 10,416.58 hours, *Semi-noisy* 2,288.06 hours, and *Noisy* 1,617.47 hours. The inclusion of noisy data captures realistic transcription variability and supports ASR training under practical conditions.

This slab-wise stratification balances dialectal coverage with data quality, enabling robust model evaluation under varying transcription conditions – essential for developing dialect-aware ASR systems.

4.2.2 Signal-Level Audio Quality

Table 4 presents signal-level audio quality metrics for *Clean*-slab across languages, including the number and percentage of low-SNR files, average words per audio, average duration, and speaking rate in words per minute (WPM). To ensure accurate measurements, each audio was trimmed using forced alignment timestamps to remove leading and trailing noise or prompts. SNR was computed using the pre-trained FB-Denoiser [34], with 4 dB empirically chosen as the threshold for classifying low-SNR files. Speaking rate was calculated as the ratio of transcript word count to the forced-alignment-based audio duration. Overall, low-SNR files make up less than 1% of the data in all languages, confirming high acoustic quality. WPM values range from 110 to 174, with lower rates observed for Kannada and Telugu due to their agglutinative linguistic structure, which results in longer word durations.

Table 4: Audio statistics per language including low SNR and speaking rate.

LID	#Files	#Low SNR	%SNR	Wds/Aud	Dur (s)	WPM
bn	870,793	3712	0.43	9	4.18	142.00
bh	866,619	4404	0.51	10	3.94	159.37
ch	823,803	1605	0.19	12	4.87	161.18
hi	756,886	1686	0.22	11	3.81	173.91
kn	744,617	1749	0.23	8	4.84	110.16
mg	968,365	2981	0.31	10	4.25	153.97
mt	518,504	1144	0.22	10	3.87	150.73
mr	1,002,599	2055	0.20	8	4.27	132.66
te	895,131	3051	0.34	8	4.40	117.16

Abbreviations: LID = Language ID; #Files = No. of audio files; #Low SNR = No. of low-SNR files (SNR < 4 dB); %SNR = Percentage of low-SNR files; Wds/Aud = Avg. words per audio; Dur (s) = Avg. duration in seconds; WPM = Words per minute.

Table 5: Train, development, and test set statistics for each language.

LID	#Dialects	Train Set				Dev Set				Test Set			
		Dur (h)	#Utts	#Sents	#Spks	Dur (h)	#Utts	#Sents	#Spks	Dur (h)	#Utts	#Sents	#Spks
bh	3	142.98	95280	19056	1445	2.14	1500	575	60	3.10	2220	694	120
bn	5	142.96	85800	17160	1280	2.27	1500	494	100	3.26	2174	648	200
ch	4	175.22	85800	17160	1586	2.40	1413	511	80	3.85	2234	695	160
hi	5	128.47	85800	17160	2172	2.21	1539	722	100	3.30	2288	853	201
kn	5	164.83	85800	17160	1859	2.37	1430	518	100	3.61	2161	663	200
mg	4	157.77	95280	19056	1493	2.10	1431	494	80	3.17	2193	640	160
mt	4	159.32	95280	19056	1913	2.06	1409	693	80	3.33	2172	993	160
mr	4	140.49	95280	19056	2305	1.98	1386	509	80	3.04	2170	711	160
te	4	155.89	95280	19056	1848	2.30	1438	500	80	3.37	2226	652	160

LID: Language ID, **#Dialects**: number of dialects, **Dur**: duration in hours, **#Utts**: number of utterances, **#Sents**: number of unique sentences, **#Spks**: number of speakers.

4.2.3 Speaker Metadata Validation

To assess the correctness and consistency of speaker metadata, we designed two validation checks: (1) intra-speaker and (2) inter-speaker. The intra-speaker check identifies inconsistencies within a single speakers recordings, while the inter-speaker check detects if recordings assigned to different speaker IDs may actually belong to the same individual. To address these inconsistencies, we developed a bucketization algorithm, which was evaluated on unseen data (see Appendix B for details). The method successfully resolved 99.28% of intra-speaker issues and 52.91% of inter-speaker mismatches, providing a reliable approximation of speaker identity consistency within the corpus. Following the speaker bucketization check, a subset of speakers with no intra- or inter-speaker discrepancies was identified and used to prepare the development and test sets, ensuring no speaker overlap across train, dev, and test splits.

5 Benchmarking ASR Performance

5.1 Datasets

To support reproducible research and enable fair benchmarking, we release standardized train, development, and test splits for each of the nine languages in the RESPIN corpus.[5] Table 5 summarizes split statistics, including duration, number of utterances, unique sentences, and speakers. Each language contains 35 dialects and 130175 hours of training audio comprising 85k95k utterances. The dev and test sets include 24 hours of speech each, with up to 2.2k utterances from 60200 distinct speakers. The train set shown in Table 5 refers to the *small* train set, a balanced subset of the *clean* corpus used for all ASR experiments in this paper. For `mt_D2`, where clean audio was insufficient, a small portion of semi-noisy data was included. Detailed statistics for other training variants are provided in the supplementary appendix.

All dev and test sets consist exclusively of speakers from the *uncontaminated* bucket (Section 4.2.3), ensuring high-quality evaluation without speaker overlap across splits. The splits are carefully constructed to preserve dialectal diversity, balance sentence types, and maintain speaker disjointnesspro-viding a reliable foundation for evaluating both traditional ASR systems and fine-tuned pretrained models.

[5]RESPIN-S1.0 data is available at: `https://github.com/saurabhk0317/respin_data_neurips25`

Table 6: CER and WER (%) for different models across languages. **Pretrained models** refer to models fine-tuned on publicly available data other than RESPIN. **Traditional models** are trained from scratch on RESPIN. **Fine-tuned models** are pretrained SSL or Whisper models further fine-tuned on a subset of RESPIN. For SeamlessM4T-v2-Large, **bh**, **ch**, and **mg**, and for the pretrained SSL models, **bh**, **ch**, **mg**, and **mt** are evaluated using Hindi-tuned models.

Model	CER (%)										WER (%)									
	bh	bn	ch	hi	kn	mg	mr	mt	te	avg	bh	bn	ch	hi	kn	mg	mr	mt	te	avg
Pretrained Models (fine-tuned on non-RESPIN public data)																				
SeamlessM4T-v2-Large	29.09	17.54	33.20	15.34	18.91	30.07	14.44	27.15	14.33	22.23	56.77	45.56	71.86	25.43	55.38	56.49	42.09	66.64	46.11	51.81
IndicW2V	17.08	14.27	22.77	11.02	10.37	19.64	15.09	23.30	8.61	15.80	51.61	42.83	65.98	28.34	42.37	54.32	53.91	66.10	37.82	49.25
SPRING-W2V2	15.10	12.50	20.81	8.80	11.43	16.35	7.56	20.12	6.97	13.29	41.32	25.93	55.42	22.99	44.35	42.09	34.15	53.69	36.32	39.58
SPRING-Data2Vec-AQC	15.02	11.94	21.26	7.20	10.78	15.81	7.49	19.91	6.53	12.88	42.35	23.69	56.17	20.93	42.79	42.47	33.40	53.65	33.98	38.83
Traditional Models (trained from scratch on RESPIN subset)																				
TDNN-HMM	5.67	5.22	4.45	3.25	4.88	7.69	3.30	6.53	3.94	4.99	17.57	16.87	12.69	8.72	23.01	22.33	13.40	20.13	20.81	17.28
E-Branchformer	4.95	4.33	3.63	3.52	4.62	6.68	3.19	5.75	3.97	4.52	15.21	14.96	10.59	9.94	24.50	20.38	14.48	17.95	21.64	16.63
Fine-tuned Models (fine-tuned on RESPIN subset)																				
Whisper-Tiny	9.62	11.60	7.13	9.69	12.62	13.98	9.15	10.73	11.43	10.66	27.45	32.51	20.81	21.71	48.54	36.40	30.93	31.96	41.61	32.44
Whisper-Base	7.15	7.69	5.36	5.80	8.10	10.44	6.23	7.51	7.51	7.31	22.51	24.71	16.67	15.19	36.52	30.54	24.28	24.80	32.99	25.36
Whisper-Small	7.90	5.46	3.85	4.16	6.00	7.46	3.93	5.94	6.54	5.69	19.02	18.91	12.36	11.78	29.66	23.94	16.95	20.28	27.82	20.08
IndicW2V	4.42	4.28	3.24	3.16	4.68	6.02	3.19	5.19	4.54	4.30	16.07	16.65	11.36	10.47	24.86	21.51	15.13	19.19	24.03	17.69
SPRING-W2V2	3.92	3.86	2.99	2.37	4.30	5.20	2.49	4.37	3.85	3.71	14.61	15.12	10.74	8.22	23.90	19.40	12.75	16.64	21.92	15.92
SPRING-Data2Vec-AQC	3.95	3.63	2.84	2.27	4.11	4.98	2.38	4.30	3.72	3.58	14.84	14.15	10.25	7.91	23.13	18.50	12.28	16.41	21.17	15.40

5.2 Existing ASR Models

We evaluate a variety of ASR models ranging from traditional models trained from scratch to modern self-supervised and multilingual pretrained models. These include Kaldi and ESPnet-based models trained on RESPIN data, as well as pretrained models like Whisper, IndicWav2Vec, and SPRING SSL models. Fine-tuning is performed wherever applicable to enable fair comparison across architectures and training paradigms.

5.3 Experimental Setup

All model training and fine-tuning experiments were performed on a single NVIDIA RTX 3090 GPU with 24GB memory.

Whisper Models: We fine-tune the Tiny (39M), Base (74M), and Small (244M) variants using Hugging Face checkpoints and the Trainer API with default settings. Fine-tuning is done separately for each language with early stopping based on validation WER. Language IDs are passed during decoding.

Fairseq Models: We fine-tune three SSL models on RESPIN: (i) IndicWav2Vec (trained on 40 Indian languages)[6], (ii) SPRING-Wav2Vec2 (30k hours, 24 languages), and (iii) SPRING-Data2Vec-AQC, which incorporates augmentation, quantization, and clustering[7].

ESPnet Models: We train a CTC-Attention hybrid model with an `e_branchformer` encoder (8 blocks, 256-dim hidden units) using Adam optimizer, SpecAugment, mixed-precision (AMP), and early stopping with patience 5 based on validation CER.

Kaldi Models: We train TDNN-HMM models using the standard chain recipe with 40-dim MFCCs, iVectors, speed/volume perturbation, and a tri-gram LM trained on RESPIN transcripts.

Pretrained models and training recipes for Whisper, Fairseq, and ESPnet experiments are available at: `https://github.com/labspire/respin_baselines`[8].

5.4 Results and discussion

Table 6 presents the ASR performance of various models on nine Indian languages using the RESPIN corpus. The results highlight the importance of dialectal supervision in training and fine-tuning ASR models.

Pretrained models struggle without dialectal supervision: Models pretrained on external corpora – such as SeamlessM4T-v2-Large, IndicW2V2 (PT), and SPRING-W2V2 (PT) – perform poorly across most languages. Their high WERs, exceeding 50% in some cases, reflect the lack of dialectal variation in the training data. Performance drops are especially evident for dialect-heavy languages like Bhojpuri and Chhattisgarhi.

[6] `https://github.com/AI4Bharat/IndicWav2Vec`
[7] `https://asr.iitm.ac.in/models`
[8] `https://github.com/labspire/respin_baselines`

Training from scratch on RESPIN improves performance: Traditional models such as TDNN-HMM and E-Branchformer, trained entirely on RESPIN subsets, significantly outperform the pre-trained models. E-Branchformer achieves an average WER of 16.63%, underscoring the benefits of dialect-specific supervision even without large-scale pretraining.

Whisper fine-tuning offers limited gains: While Whisper models (Tiny, Base, Small) fine-tuned on RESPIN perform better than their pretrained-only versions, they still lag behind scratch-trained models. Whisper-Small, for instance, shows higher WER than E-Branchformer despite using the same data, suggesting that general-purpose multilingual models do not fully adapt to dialectal variation.

SSL models fine-tuned on RESPIN perform best: Self-supervised models like SPRING-W2V2 and SPRING-Data2Vec-AQC, when fine-tuned on RESPIN, outperform all other approaches. SPRING-Data2Vec-AQC achieves the lowest average WER (15.40%) and delivers the best performance across most individual languages, demonstrating the strength of combining SSL pretraining with dialect-aware fine-tuning.

Summary: These findings show that RESPINs dialectal coverage provides clear benefits across model types. Pretrained models struggle with domain mismatch, while both scratch-trained and fine-tuned models gain significantly from RESPINs diversity. Fine-tuned SSL models emerge as the most effective strategy for multi-dialect ASR in the Indian context.

6 Applications, Impact, and Limitations

RESPIN-S1.0 has already made a tangible impact within the speech technology community. Over the past two years, subsets of the corpus have been released through various workshops and challenges. A subset of Bengali and Bhojpuri data was used in the SLT Code Hackathon 2022 to build dialectal ASR systems. The first Multi-Dialect ASR Challenge (MADASR) was held at ASRU 2023 [35, 36] using RESPIN data for Bengali and Bhojpuri, and the ongoing MADASR 2.0 Challenge at ASRU 2025 expands this to 1,200 hours across eight languages (bh, bn, ch, kn, mg, mr, mt, te), enabling large-scale benchmarking of dialect-aware ASR systems. RESPIN has also been used to study dialect identification performance across eight Indian languages [37]. Beyond ASR, the corpus supports a range of tasks including language and dialect identification (LID/DID), unsupervised speech translation, and broader speech-language research, particularly for underrepresented Indian languages and domain-specific applications. However, RESPIN-S1.0 has some limitations. It currently contains only read speech, while spontaneous or conversational data is often more representative of real-world scenarios. It is also limited to two domainsagriculture and financeselected for societal relevance; expanding to domains such as healthcare and education could enhance applicability. Finally, while efforts were made to ensure inclusivity, the reliance on literate native speakers with mobile access may underrepresent the most marginalized populations. Nevertheless, RESPIN sets a strong foundation for inclusive and dialect-rich ASR development in India, and future versions will address these limitations through broader linguistic coverage and inclusion of spontaneous speech.

7 Conclusion and Future Work

RESPIN-S1.0 is the first large-scale, publicly available corpus that combines dialectal and domain coverage across nine Indian languages, including low-resource ones like Bhojpuri, Chhattisgarhi, and Magahi. By addressing long-standing gaps in linguistic diversity, speaker variation, and domain relevance, RESPIN enables the development of more inclusive and robust speech technologies for Indian languages. The standardized benchmarks, metadata, and carefully curated splits provided with this release support reproducible research in ASR and beyond. As part of ongoing efforts, RESPIN-S2.0 will expand to cover additional languages, dialects, spontaneous speech, and new domains. We invite researchers, institutions, and industry collaborators to join us in building the next generation of speech and language tools that reflect the full diversity of India's linguistic landscape.

Acknowledgements

This work was supported by the Bill and Melinda Gates Foundation. We thank the many NGOs, volunteers, and contributors who participated in data collection, validation, and community engagement – both online and offline. Their efforts were instrumental in making RESPIN-S1.0 possible.

References

[1] Nandini Sethi and Amita Dev. Survey on automatic speech recognition systems for indic languages. In Amita Dev, S. S. Agrawal, and Arun Sharma, editors, *Artificial Intelligence and Speech Technology*, pages 85–98, Cham, 2022. Springer International Publishing.

[2] John J Gumperz. Speech variation and the study of indian civilization. *American Anthropologist*, 63(5):976–988, 1961.

[3] World Urbanization Prospects World Bank staff estimates based on the United Nations Population Divisions. Rural population (% of total population) - india. $https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=IN$, Last accessed 31 July 2023.

[4] Hari Krishna Vydana. *Salient Features for Multilingual Speech Recognition in Indian Scenario*. PhD thesis, Jan 2019. $http://hdl.handle.net/10603/246056$.

[5] R. B. Mandal. Patterns of regional geography:an international perspective ů volume 2. page 140, 1990.

[6] K. Samudravijaya. Indian Language Speech Label (ILSL): A De Facto National Standard. In Anupam Biswas, Emile Wennekes, Tzung-Pei Hong, and Alicja Wieczorkowska, editors, *Advances in Speech and Music Technology*, pages 449–460, Singapore, 2021. Springer Singapore.

[7] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Interspeech 2018*, pages 3743–3747, 2018.

[8] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. E-Branchformer: Branchformer with Enhanced Merging for Speech Recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 84–91, 2023.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[10] Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Towards Building ASR Systems for the Next Billion Users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[11] Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, et al. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *arXiv preprint arXiv:2403.01926*, 2024.

[12] Ashwin Sankar, Srija Anand, Praveen Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *Advances in Neural Information Processing Systems*, 37:68161–68182, 2024.

[13] Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12942–12950, 2023.

[14] Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 1–5. IEEE, 2023.

[15] Kaushal Santosh Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, and Pratyush Kumar. Vistaar: Diverse benchmarks and training sets for indian language asr. *arXiv preprint arXiv:2305.15386*, 2023.

[16] Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M Khapra. Svarah: Evaluating english asr systems on indian accents. *arXiv preprint arXiv:2305.15760*, 2023.

[17] Arjun Gangwar, S Umesh, Rithik Sarab, Akhilesh Kumar Dubey, Govind Divakaran, Suryakanth V Gangashetty, et al. Spring-inx: A multilingual indian language speech corpus by spring lab, iit madras. *arXiv preprint arXiv:2310.14654*, 2023.

[18] Abhayjeet Singh, Charu Shah, Rajashri Varadaraj, Sonakshi Chauhan, and Prasanta Kumar Ghosh. Spire-sies: A spontaneous indian english speech corpus. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE, 2023.

[19] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[20] Anish Bhanushali, Grant Bridgman, Prasanta Ghosh, Pratik Kumar, Saurabh Kumar, Adithya Raj Kolladath, Nithya Ravi, Aaditeshwar Seth, Ashish Seth, Abhayjeet Singh, et al. Gram vaani asr challenge on spontaneous telephone speech recordings in regional variations of hindi. In *Proc. Interspeech 2022*, pages 3548–3552, 2022.

[21] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, et al. Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*, 2021.

[22] Devaraja Adiga, Rishabh Kumar, Amrith Krishna, Preethi Jyothi, Ganesh Ramakrishnan, and Pawan Goyal. Automatic speech recognition in sanskrit: A new speech corpus and modelling insights. *arXiv preprint arXiv:2106.05852*, 2021.

[23] Joyanta Basu, Soma Khan, Rajib Roy, Tapan Kumar Basu, and Swanirbhar Majumder. Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification. *Circuits, Systems, and Signal Processing*, 40(10):4986–5013, 2021.

[24] Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Prashant Krishnan, et al. Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957. IEEE, 2021.

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[26] Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, et al. Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, 2020.

[27] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, 2020.

[28] Nimisha Srivastava, Rudrabha Mukhopadhyay, CV Jawahar, et al. Indicspeech: Text-to-speech corpus for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6417–6422, 2020.

[29] Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In *SLTU*, pages 11–14, 2018.

[30] Keshan Sodimana, Pasindu De Silva, Supheakmungkol Sarin, Oddur Kjartansson, Martin Jansche, Knot Pipatsrisawat, and Linne Ha. A step-by-step process for building tts voices using open source data and frameworks for bangla, javanese, khmer, nepali, sinhala, and sundanese. In *SLTU*, pages 66–70, 2018.

[31] Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, and Anil Kumar Vuppala. Iiith-ilsc speech database for indain language identification. In *SLTU*, pages 56–60, 2018.

[32] Andrew Wilkinson, Alok Parlikar, Sunayana Sitaram, Tim White, Alan W Black, and Suresh Bazaj. Open-source consumer-grade indic text to speech. In *SSW*, pages 190–195, 2016.

[33] Kishore Prahallad, Naresh Kumar Elluru, Venkatesh Keri, S Rajendran, and Alan W Black. The iiit-h indic speech databases. In *Interspeech*, pages 2546–2549, 2012.

[34] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

[35] Sathvik Udupa, Jesuraja Bandekar, G Deekshitha, Saurabh Kumar, Prasanta Kumar Ghosh, Sandhya Badiger, Abhayjeet Singh, Savitha Murthy, Priyanka Pai, Srinivasa Raghavan, and Raoul Nanavati. Gated multi encoders and multitask objectives for dialectal speech recognition in indian languages. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8, 2023.

[36] Tanel Alumäe, Jiaming Kong, and Daniil Robnikov. Dialect Adaptation and Data Augmentation for Low-Resource ASR: Taltech Systems for the Madasr 2023 Challenge. In *Proc. ASRU*, pages 1–7, 2023.

[37] Amartyaveer, Saurabh Kumar, Sumit Sharma, Sathvik Udupa, Sandhya Badiger, Abhayjeet Singh, Deekshitha G, Jesuraja Bandekar, Savitha Murthy, and Prasanta Kumar Ghosh. Improving dialect identification in indian languages using multimodal features from dialect informed asr. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

A Text Data Creation and Validation Details

A.1 Detailed Text Validation Checks

The following checks were implemented as part of the text validation pipeline. These include a combination of automatic and manual steps designed to ensure that the composed text adheres to the required linguistic, formatting, and usability standards across all dialects.

1. **Duplicate Sentence Removal (Automatic)**
   A pairwise Word Error Rate (WER) analysis is applied to the raw sentence set to identify and remove duplicates. Performing this step early reduces unnecessary overhead for downstream validators.
2. **Invalid Character Check and Correction (Manual)**
   This step eliminates non-printable characters, newline characters, and redundant whitespace. A list of all characters present in the corpus is generated and provided to validators. Sentences containing non-language characters (i.e., anything other than alphabets, numerals, and allowed punctuation: comma, full stop, and question mark) are flagged and corrected. This process is iterated until the corpus contains only valid characters.
3. **Sentence Pruning to Specified Length (Manual)**
   Due to constraints in the recording application, sentence length was capped at 90 characters. Sentences exceeding this threshold were manually pruned or rejected by validators.
4. **Acronym Standardization (Manual)**
   Acronyms are required to follow a standard "A.B.C." format, where A, B, and C are characters in the acronym. Words containing full stops are extracted and validated to ensure they are either valid acronyms or corrected appropriately. Sentences containing unformatted acronymsidentified via transliterated English tokens or known acronym listsare also flagged and corrected.
5. **Invalid Matra Check and Correction (Manual)**
   Words with incorrect or redundant matra usage (e.g., consecutive matras or visually overlapping matras with identical appearance) are flagged and manually corrected to maintain script correctness.
6. **Interchangeable Character Word Correction (Manual)**
   Validators provide a list of commonly confused or interchangeable characters. Sentences containing words with these characters are reviewed for potential spelling errors and corrected accordingly.
7. **Similar Sentence Check (Manual)**
   This check builds on duplicate removal by identifying sentence pairs with $0 < \text{WER} < 0.3$. These near-duplicate pairs are reviewed by validators, who decide whether to retain, correct, or reject one of the variants.
8. **Homophone Check (Manual)**
   Using phonetic transcriptions provided by Navana Tech, phonetic WER is computed across word pairs to identify homophones. Validators assess these pairs and flag incorrect spellings to ensure consistency and correctness in pronunciation-sensitive cases.
9. **Language-Specific Checks (Manual)**
   While the above checks cover the majority of validation needs, additional checks were applied in select languages to address script-specific or dialect-specific issues. Details of these language-level customizations are provided in respective language sections of the corpus documentation.

B Speaker ID Bucketization Procedure

In crowd-sourced audio collection settings, accurately capturing speaker identity metadata is particularly challenging. Errors in speaker IDs can be categorized into two types:

- **Intra-speaker errors:** Cases where a single speaker's data is incorrectly tagged under multiple IDs.

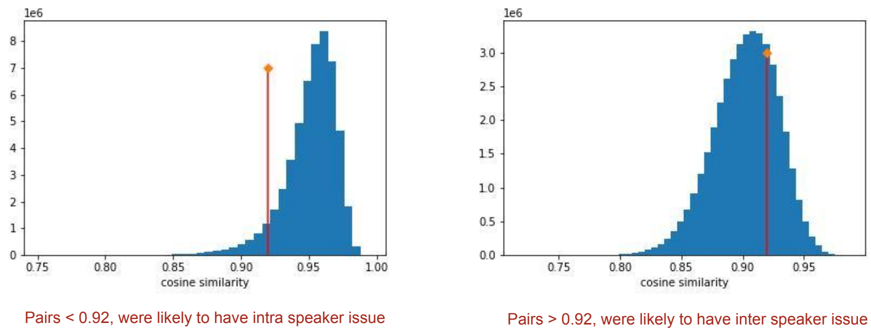

Figure 5: Cosine similarity distributions for speaker embedding pairs: (left) intra-speaker comparisons showing errors for similarities < 0.92, and (right) inter-speaker comparisons showing errors for similarities > 0.92.

- **Inter-speaker errors:** Cases where multiple speakers are erroneously tagged with the same ID.

To identify these inconsistencies, we extract speaker embeddings using a pre-trained x-vector TDNN model from SpeechBrain[9]. Cosine similarities are then computed between speaker embedding pairs:

- For **intra-speaker validation**, we compute pairwise cosine similarity between recordings assigned to the same speaker ID.
- For **inter-speaker validation**, we compute similarity between embeddings from different speaker IDs within the same district.

Figure 5 illustrates these distributions for Bengali: the left panel shows intra-speaker similarity, and the right panel shows inter-speaker similarity.

To establish thresholds, we manually inspect audio pairs sampled across cosine similarity bins (step size: 0.01). In the case of intra-speaker validation, samples are reviewed in descending order of similarity, while inter-speaker pairs are reviewed in ascending order. We empirically determine a threshold of 0.92 for both casesbelow this value, intra-speaker mismatches are likely; above this, inter-speaker identity collisions occur.

This analysis allows us to flag "contaminated" speaker IDs and curate an *uncontaminated* speaker subset. These uncontaminated speakers are subsequently used for development and test set preparation to ensure no overlap with training speakers and maintain evaluation integrity.

[9] `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is a database paper and we do not claim any theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper reports standard ASR evaluation metrics (CER and WER) to benchmark model performance on RESPIN-S1.0. Since the focus is on dataset release and not on statistically comparing methods across multiple runs or random seeds, statistical significance testing or error bars were not applicable.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The RESPIN-S1.0 dataset consists of curated, read speech collected with informed consent for research use in Indian languages. It does not contain personally identifiable information or content with high risk for misuse. Hence, specific safeguards were not required beyond standard ethical data collection practices.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The RESPIN-S1.0 corpus introduces new language and dialect-level speech and text resources for 9 Indian languages. Detailed documentation is provided alongside the assets, including data format descriptions, speaker metadata, train/dev/test splits, validation procedures, and usage instructions, hosted at `https://github.com/saurabhk0317/respin_data_neurips25`.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The RESPIN corpus was created using contributions from trained dialect experts and volunteers. While informed consent was obtained and contributors were compensated, the paper does not include the full set of instructions, screenshots, or detailed compensation information.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: All participants contributed voluntarily with informed consent for the public release of their anonymized data. The project received ethics clearance through an internal review process at IISc Bangalore, and no personally identifiable information (PII) was collected. The risks to participants were minimal and clearly communicated.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were used only for improving the writing and editing of the manuscript. They were not involved in data creation, modeling, evaluation, or any part of the research methodology.