Task-Lens: Cross-Task Utility Based Speech Dataset Profiling for Low-Resource Indian Languages

Swati Sharma, Divya V. Sharma, Anubha Gupta
SBILab, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
Delhi, India
{swati21568, divyas, anubha}@iiitd.ac.in

Abstract

The rising demand for inclusive speech technologies amplifies the need for multilingual datasets for Natural Language Processing (NLP) research. However, limited awareness of existing task-specific resources in low-resource languages hinders research. This challenge is especially acute in linguistically diverse countries, such as India. Cross-task profiling of existing Indian speech datasets can alleviate the data scarcity challenge. This involves investigating the utility of datasets across multiple downstream tasks rather than focusing on a single task. Prior surveys typically catalogue datasets for a single task, leaving comprehensive cross-task profiling as an open opportunity. Therefore, we propose Task-Lens, a cross-task survey that assesses the readiness of 50 Indian speech datasets spanning 26 languages for nine downstream speech tasks. First, we analyze which datasets contain metadata and properties suitable for specific tasks. Next, we propose task-aligned enhancements to unlock datasets to their full downstream potential. Finally, we identify tasks and Indian languages that are critically underserved by current resources. Our findings reveal that many Indian speech datasets contain untapped metadata that can support multiple downstream tasks. By uncovering cross-task linkages and gaps, Task-Lens enables researchers to explore the broader applicability of existing datasets and to prioritize dataset creation for underserved tasks and languages.

Keywords: Indian speech datasets, multilingual resources, low-resource languages, cross-task profiling, dataset readiness, metadata analysis, speech resource survey, dataset reuse, data scarcity

1. Introduction

The growing demand for inclusive speech technologies intensifies the need for multilingual datasets for Natural Language Processing (NLP) research. However, most speech datasets are English-centric, so the lack of speech datasets in low-resource languages hinders NLP research. Multilingual speech models aim for inclusivity but often underperform and exhibit linguistic biases in various tasks (Xu et al., 2020). Consequently, there is an urgent need for task-specific datasets for low-resource languages. This data scarcity challenge can be alleviated by efficient utilization of existing resources. However, researchers are often unaware of existing resources for underrepresented languages, which hinders NLP research for those languages (Larasati, 2025). Cross-task profiling is a viable solution to address this problem. Cross-task profiling involves the systematic analysis of dataset attributes to assess their readiness for multiple downstream tasks beyond their originally intended use.

Cross-task profiling can reveal how existing datasets support diverse tasks. However, most prior works describe dataset creation or catalogue Indian speech resources without profiling them across tasks (Shrishrimal et al., 2012; Kurian, 2015; Petkar, 2017; Palia et al., 2013; Verma et al., 2018; Singh et al., 2020). Recent South Asian surveys present NLP progress (data, models, tasks) across languages with only incidental coverage of speech, and primarily catalogue speech resources without any cross-task profiling (Poria and Huang, 2025). As a result, the community often lacks clarity on whether existing speech resources are usable across tasks, especially for underrepresented languages. Although these challenges are global, they are particularly acute in linguistically diverse settings such as India. Cross-task profiling of Indian speech datasets can help researchers maximise the value of available resources and motivate data collection efforts for low-resource languages, which is essential for advancing multilingual speech research.

In this work, we present **Task-Lens**, a cross-task survey of 50 Indian speech datasets. Beyond cataloging datasets, Task-Lens brings to light resources with utility across multiple tasks that typical AI tools or web searches cannot reliably surface. It highlights Indian speech resources, which include task-aligned metadata that can enable cross-task reuse. Thus, providing readiness profiles that help researchers quickly discover corpora for specific tasks. It also identifies actionable gaps, including missing features that limit cross-task readiness, and pinpoints critically underserved languages and tasks to motivate targeted dataset creation. By reducing discovery time and clarifying reuse pathways, Task-Lens turns scattered resources into a navigable map for advancing multilingual speech research in India.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5883–5895
11–16 May 2026.

We summarize our main contributions below:

1. We present **Task-Lens**, a cross-task survey that profiles the readiness of Indian speech datasets for multiple downstream tasks using available metadata.
2. Task-Lens includes comprehensive cross-task profiling across nine downstream tasks, using 50 Indian speech datasets covering 26 languages and comprising over 91,257 hours of audio.
3. We investigate the following research questions: (a) Which speech tasks does each dataset currently support? (b) What improvements would enhance a dataset's suitability for additional tasks? (c) Which areas of speech research lack adequate dataset support? (d) Which Indian languages offer adequate coverage per task, and where do significant resource gaps remain?

2. Related Works

Cross-Task Utility Exploration: Previous works explore profiling frameworks that measure representations across global languages and tasks (Yang et al., 2021; Conneau et al., 2022; Gebru et al., 2021; Chen et al., 2024; Mazumder et al., 2023). However, most surveys and benchmarks have focused on cataloging available datasets (Poria and Huang, 2025; Shilin et al., 2018; Bakhturina et al., 2021). Thus, several key research questions are underexplored in the literature, such as: (1) While resource papers often associate datasets with specific tasks, rich metadata makes them useful for broader applications. How can we investigate the cross-task utility of existing multilingual speech resources? (2) Is there any urgent need for speech datasets of some specific language for specific tasks? With over 7,000 languages spoken worldwide, addressing these questions can uncover key NLP research gaps. This is especially important for linguistically diverse countries such as India. This study contributes a cross-task resource analysis that evaluates the readiness of Indian speech datasets for diverse downstream tasks using their documented metadata and properties.

Indian Dataset Landscape: India's 22 official languages remain underrepresented in mainstream speech benchmarks, and high-quality, general-purpose corpora are scarce (Joshi et al., 2020). Early surveys cataloged Indian speech resources and noted missing standardization, scale, and metadata consistency (Shrishrimal et al., 2012). Task-focused reviews address specific areas for Automatic Speech Recognition, Language Identification, Code-switching, Text-to-Speech, and Speech Emotion Recognition (Singh et al., 2020; Unnibhavi and Jangamshetti, 2016; Bakshi and Kopparapu, 2018; Dey et al., 2023; Sitaram et al., 2019; Mustafa et al., 2022; Agro et al., 2025; Panda et al., 2020; Monisha and Sultana, 2022). While these efforts advance the coverage of techniques and challenges, the cross-task profiling of existing speech datasets continues to represent an open research opportunity. Furthermore, the closest prior studies to ours are Shrishrimal et al. (2012) and Poria and Huang (2025). Shrishrimal et al. (2012) catalogs Indian speech resources but provides no guidance on repurposing metadata beyond their original intent. Moreover, it includes no resources published after 2012. By contrast, Poria and Huang (2025) presents a timely, well-curated survey of South Asian NLP since 2020, with emphasis on data, models, and tasks. However, this paper is not Indian-speech focused, treats speech only incidentally, and does not articulate speech-specific task-readiness criteria or cross-task reuse pathways. In this work, we move beyond dataset cataloging to perform cross-task profiling of 50 Indian speech datasets covering 26 low-resource Indian languages over nine downstream tasks. Our goal is to highlight available datasets that include sufficient metadata to support tasks beyond their originally intended purpose. This cross-task perspective enables researchers working on underserved languages and speech processing tasks to identify datasets with potential utility for their research. Furthermore, our analysis reveals significant gaps across specific languages and tasks, thereby motivating future data collection efforts targeting these underrepresented areas.

3. Task-Lens

Task-Lens is a cross-task, utility oriented lens for profiling Indian language speech datasets. The pipeline comprises four stages: dataset discovery, dataset filtering, feature extraction, and utility mapping, as shown in Figure 1. This section reports the methodology in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for transparent and reproducible evidence synthesis (Page et al., 2021). In line with PRISMA, we detail the eligibility criteria, information sources, search strategy, selection process, data items, etc.

Dataset Discovery: We searched peer-reviewed venues and dataset registries relevant to speech and language resources in India. Sources included IEEE Xplore, ACM Digital Library, ACL Anthology, ISCA Interspeech, LREC ELRA, Scopus, arXiv, etc.; portals included OpenSLR, Bhashini ULCA and Vatika, AIKosh, LDC IL, ELRA Catalogue, Mozilla Common Voice, Google FLEURS, AI4Bharat repositories, Hugging Face Datasets, Zenodo, and GitHub releases.

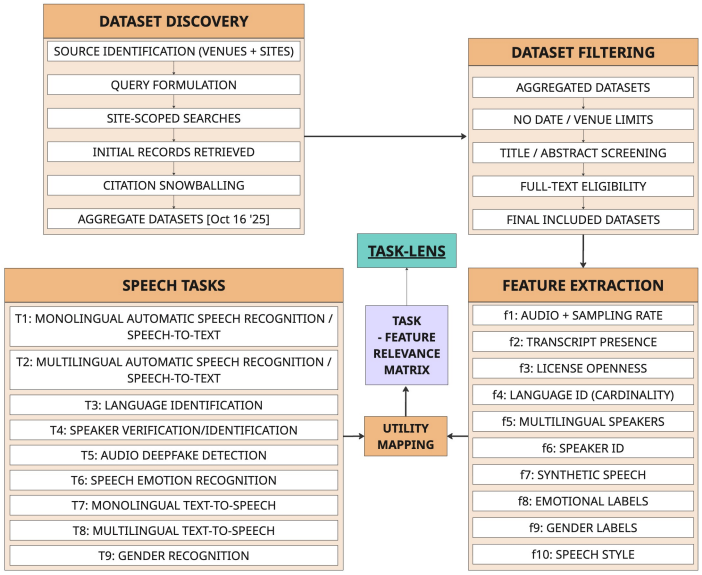

Figure 1: Task-Lens: It involves dataset discovery, dataset filtering, feature extraction, followed by utility mapping that aligns dataset features with task needs via a Task-feature relevance matrix labeled as Required and Optional or Not Applicable. A dataset is 'Task-Ready' for a task if it satisfies all 'Required' features for a task. Supported tasks include Automatic Speech Recognition (Monolingual) [T1], Automatic Speech Recognition (Multilingual) [T2], Language Identification [T3], Speaker Verification/Identification [T4], Audio Deepfake Detection [T5], Speech Emotion Recognition [T6], Text-to-Speech (Monolingual) [T7], Text-to-Speech (Multilingual) [T8] and Gender Recognition [T9].

We used queries that combine resource, task, and language terms (e.g., Indian/Indic/Hindi/Tamil; "speech dataset"/"speech corpus"; and "automatic speech recognition"/"language identification"/"text-to-speech"/"speech emotion recognition"/ "gender recognition"/"audio deepfake detection") and issued site-scoped searches for each portal. We imposed no date limits or venue filters. We also applied backward and forward citation snowballing from seed papers and dataset pages. The dataset search has been ongoing for several months, with the last comprehensive search conducted on October 16, 2025.

Dataset Filtering: We applied predefined eligibility criteria in two stages: title and abstract screening, followed by full text or webpage review. The first stage removed clear textual records, non-Indic resources, and duplicates. The second stage confirmed the presence of basic metadata, including audio files, sampling rate, and practical accessibility. Inclusion required Indian languages, public documentation with extractable features, and identifiable splits or benchmark usage. Synthetic speech was conditionally included when provenance and generation were documented.

Following this procedure, we curated a list of 50 Indian speech datasets. The complete list of datasets, along with their abbreviations and references, is presented in Table 1. These datasets cover speech in 26 languages spoken across the Indian demographic population, collectively providing over 91,257 hours of audio. The collection comprises monolingual and multilingual datasets, covering a diverse range of speech styles, domains, speaker types, and annotation depths.

Feature Extraction: After filtering, we extracted 10 descriptive features for each dataset using a standardized schema. Table 2 summarizes the selected features and their rationale. Each feature will be referenced as presented in the table, in the form f_i[1].

Utility Mapping: After the feature extraction stage, we get a set of features for each dataset. The next step is to perform a task utility mapping for these datasets. Each task will be referenced in the form T_j. For utility mapping, we considered nine representative speech-technology tasks:

T_1 : Monolingual Automatic Speech Recognition / Speech-to-Text (MO-ASR/STT)

T_2 : Multilingual Automatic Speech Recognition / Speech-to-Text (ML-ASR/STT)

T_3 : Language Identification (LID)

T_4 : Speaker Verification/Identification (SV/SID)

T_5 : Audio Deepfake Detection (ADD)

T_6 : Speech Emotion Recognition (SER)

[1]All per-dataset feature values have been released as supplementary material.

ID	Dataset
D_1	AccentDB Ahamad et al. (2020)
D_2	Assamese TTS Corpus Tamim et al. (2025)
D_3	BhasaAnuvaad Jain et al. (2024)
D_4	Open-source TTS Voices Sodimana et al. (2018)
D_5	Bengali Numbers Corpus Nahid et al. (2018)
D_6	South Asian Crowdsourced Speech Kjartansson et al. (2018)
D_7	Low-Income Workers Abraham et al. (2020)
D_8	FLEURS Conneau et al. (2022)
D_9	GACMIS Songs Ujjwal et al. (2020)
D_{10}	GlobalPhone Speaker Package ELRA (2018)
D_{11}	GRAM VAANI Bhanushali et al. (2022)
D_{12}	Hindi-Tamil-English ASR IIT-M (2021)
D_{13}	Indian Folk Music Singh et al. (2022)
D_{14}	Regional Music Singh and Biswas (2021)
D_{15}	Indic TTS (IITM) IIT Madras (2025)
D_{16}	IndicSpeech (TTS) Srivastava et al. (2020)
D_{17}	IndicSUPERB Javed et al. (2023b)
D_{18}	IndicVoices-R Sankar et al. (2024)
D_{19}	Kashmiri Data Corpus OpenSLR
D_{20}	KritiSamhita Konduri et al. (2024)
D_{21}	Lahaja Javed et al. (2024)
D_{22}	MS Indic Speech Microsoft (2024)
D_{23}	MUCS Site Diwan et al. (2021)
D_{24}	Nexdata AI 759h Nexdata AI
D_{25}	NISP Kalluri et al. (2021)
D_{26}	NPTEL2020 AI4Bharat (2020)
D_{27}	Opensource Multispeaker Data He et al. (2020)
D_{28}	Rajasthani Hindi (MS) Microsoft (2024)
D_{29}	Rasa Srinivasa Varadhan et al. (2024)
D_{30}	SMC Malayalam Computing (2020)
D_{31}	Svarah Javed et al. (2023a)
D_{32}	Urdu Recognition (Desktop) ELRA (2024)
D_{33}	Vāksañcayaḥ Adiga et al. (2021)
D_{34}	IndicSynth Sharma et al. (2025)
D_{35}	NIRANTAR Kumar et al. (2025b)
D_{36}	Shrutilipi-Anuvaad Pothula et al. (2025)
D_{37}	IIITH-HE-CM Rambabu and Gangashetty (2018)
D_{38}	EmoTa: A Tamil Emotional Speech Dataset Thevakumar et al. (2025)
D_{39}	IndicFake Ranjan et al. (2025)
D_{40}	SPIRE-SIES Prabhu et al. (2023)
D_{41}	BanglaSER Das et al. (2022)
D_{42}	SUBESCO Sultana et al. (2021)
D_{43}	KBES Billah et al. (2023)
D_{44}	BanSpEmo Sultana et al. (2025)
D_{45}	SEA_Spoof Wu et al. (2025)
D_{46}	IIITH-ILSC Vuddagiri et al. (2018)
D_{47}	I-MSV Mishra et al. (2023)
D_{48}	IndieFake Kumar et al. (2025a)
D_{49}	Bangla Speech Corpus Ahmed et al. (2020)
D_{50}	Indic-TEDST Sethiya et al. (2024)

Table 1: List of the 50 Indian speech datasets included in our analysis. Each dataset is assigned a unique identifier (D_1–D_{50}), which is used consistently throughout the paper. The table includes established and lesser-known resources, along with relevant citations.

T_7 : Monolingual Text-to-Speech (MO-TTS)

T_8 : Multilingual Text-to-Speech (ML-TTS)

T_9 : Gender Recognition (GRE)

Here, T_1 - T_2 and T_7 - T_8 are treated separately to reflect the difference between generating text/speech across multiple languages and producing high-quality output in a single language.

Task–Feature Mapping: The task–feature relevance matrix in Table 3 reflects a rule-based mapping grounded in task literature and benchmark specifications. For each pair (f_i, T_j), we first formalized the task and its core data needs. For example, LID prioritizes utterance level language labels rather than transcripts. We then validated requirements using benchmark documentation and recent surveys across Indian speech research (Panda et al., 2020; Monisha and Sultana, 2022; Kurian, 2015; Bakshi and Kopparapu, 2018; Singh et al., 2020; Poria and Huang, 2025; Dey et al., 2023). We minimized required features to ensure consistent and fair utility judgments across datasets and tasks. The mapping remains extensible as task definitions evolve or new task families emerge.

For each task t and feature f, we assign a categorical label $\{\checkmark, ?\}$, where $\checkmark$ denotes Required, ? denotes Optional or Not Required. Table 3 summarizes these labels for all combinations of tasks and features.

Feature	Rationale
f_1: Audio + Sampling Rate	Ensures audio fidelity and robustness
f_2: Transcript Presence	Enables text-speech alignment
f_3: License Openness	Governs accessibility reuse
f_4: Language ID (Cardinality)	Supports multilingual tasks
f_5: Multilingual Speakers	Facilitates code-switching research
f_6: Speaker ID	Essential for speaker-centric tasks
f_7: Synthetic Speech	Used for augmentation/detection
f_8: Emotional Labels	Supports affective computing
f_9: Gender Labels	Enables fairness/robustness analyses
f_{10}: Speech Style	Distinguishes read vs. conversational/scripted, affecting task transferability

Table 2: Utility feature summary.

4. Task-Lens Utility Exploration

4.1. Cross-Task Dataset Utility

Standard practice in speech processing involves designing each dataset d for a single task t. However, published datasets often contain rich metadata, making them suitable for other tasks too. For instance, Librispeech was introduced for the ASR

task but has been leveraged for speaker verification task (Panayotov et al., 2015; Sharma and Buduru, 2022). The cross-task utility of published datasets is an area that remains underexplored. Consequently, due to a lack of cross-task utility exploration, published datasets are often underutilized for diverse applications despite their potential utility.

The lack of cross-task studies creates additional challenges for NLP researchers working on low-resource language tasks. Therefore, we use Task-Lens to address the following research questions: *(1) Which tasks t does each dataset d currently support? (2) What enhancements would make a dataset suitable for cross-task applications?*

Setup. For each dataset d and task t, Task-Lens checks whether all features marked ✓ (Required) in the task-feature relevance matrix (Table 3) are present. A dataset is 'Task-Ready' for a task if it satisfies all 'Required' features for a task. Task and dataset details appear in Section 3 and Table 1.

Observations: As shown in Table 4, the corpus set demonstrates broad task coverage across datasets. We observed that datasets D_4, D_6, D_{15}, D_{16}, D_{18}, D_{22}, D_{29}, D_{34}, and D_{35} contain the required features to support seven of the nine tasks. These datasets commonly lack speaker identifiers (f_6), synthetic speech (f_7), or emotion labels (f_8), which are essential for extending usability to speaker verification (T_4), deepfake detection (T_5), and emotion recognition (T_6). All of them would reach complete coverage by incorporation of the above. Furthermore, multiple *Task-Ready* datasets qualify for specific objectives and fall short of only a few features in supporting additional tasks. From a development perspective, several datasets represent clear candidates for improvement, where targeted inclusion of these key metadata would substantially enhance their cross-task readiness.

4.2. Task-Wise Data Requirement

Developing robust speech models for Indian languages requires a clear understanding of the existing datasets for each task and identifying where gaps persist. Although ASR, TTS, and GRE resources have expanded rapidly (Table 4), the distribution of datasets across tasks remains uneven and opaque to practitioners (Javed et al., 2023b; Sankar et al., 2024). Researchers often spend significant effort surveying repositories to determine whether tasks such as emotion recognition or deepfake detection have sufficient data, but find limited or no Indian-specific resources (Busso et al., 2008). These observations motivate our second inquiry: *Which tasks lack sufficient dataset support for the Indian population?*

Setup: We used Task-Lens outputs (Table 4) to select tasks for each dataset. Next, we compiled, for each task, a list of datasets that satisfy it, yielding multiple Task-Ready datasets per task. After identifying these datasets, we examined their total speech duration (Figure 2) to assess coverage.

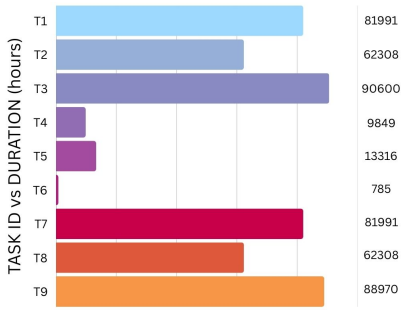

Figure 2: Distribution of total dataset duration for each task in hours for direct comparison. There is an urgent need of datasets for tasks T_4 (SV/SID), T_5 (ADD), and T_6 (SER).

Observations: Figure 2 shows that tasks T_3 (LID) and T_9 (GRE) achieve the highest coverage at roughly 90,000 hours, reflecting support from multiple large scale datasets and feature requirements that are common across many corpora. Tasks T_1 (MO-ASR/STT), T_2 (ML-ASR/STT), T_7 (MO-TTS) and T_8 (ML-TTS) each reach about 60,000 hours, indicating reasonable, though still improvable, support driven by overlapping feature needs. In contrast, tasks T_4 (SV/SID) and T_5 (ADD) cover only about 9,000 to 13,000 hours, underscoring the frequent absence of speaker identifiers and synthetic speech in low resource datasets. Finally, there is an urgent need for emotion labels in Indian language speech corpora, as only 785 hours currently support SER. Some examples of datasets with highest duration for T_1 and T_7 are: D_3, D_{26}, D_{34}, D_{35}, D_{18}; for T_2 and T_8: D_3, D_{34}, D_{35}, D_{18}, D_{17}; for T_3 and T_9: D_3, D_{26}, D_{39}, D_{34}, D_{35}; for T_4: D_{35}, D_{18}, D_6, D_{22}, D_{24}; for T_5: D_{39}, D_{34}, D_{49}, D_{45}, D_{48}; and for T_6 are: D_{29}, D_{41}, D_{42}, D_{44}, D_{38}.

4.3. Linguistic Data Requirement

Researchers in the NLP community strive to develop multilingual and inclusive speech technologies. However, the majority of speech datasets are in English. Consequently, the lack of low-resource language datasets hinders the development of multilingual speech technologies. Only a few of the world's languages (approx 7,000) have sufficient resources for human language technologies (Besacier et al., 2014). This gap motivates our third inquiry: *Which Indian languages have adequate dataset support for each task, and where do critical language-specific gaps persist?*

Setup: For each task, we selected all the Task-Ready datasets listed in Table 4. Next, we identified

Feature (f_i)	MO-ASR/ STT	ML-ASR/ STT	LID	SV/SID	ADD	SER	MO-TTS	ML-TTS	GRE
f_1: Audio + Sampling Rate	✓	✓	✓	✓	✓	✓	✓	✓	✓
f_2: Transcript Presence	✓	✓	?	?	?	?	✓	✓	?
f_3: License Openness	?	?	?	?	?	?	?	?	?
f_4: Language ID (Cardinality)	?	✓	✓	?	?	?	?	✓	?
f_5: Multilingual Speakers	?	✓	?	?	?	?	?	✓	?
f_6: Speaker ID	?	?	?	✓	?	?	?	?	?
f_7: Synthetic Speech	?	?	?	?	✓	?	?	?	?
f_8: Emotional Labels	?	?	?	?	?	✓	?	?	?
f_9: Gender Labels	?	?	?	?	?	?	?	?	✓
f_{10}: Speech Style	?	?	?	?	?	?	?	?	?

Table 3: Task–feature relevance matrix for dataset–task screening. Symbols: ✓ = Required, ? = Optional / Not Required. For each task, entries marked ✓ form the minimal, exhaustive set of required features under standard task definitions.

the languages present in those datasets, encompassing a total of *26 languages*. Next, we calculated total audio duration of each language across these datasets. Table 5 shows the total speech durations (in hours) for Task-Ready datasets across 26 languages and nine downstream tasks (T_1–T_9). The notation for each language (L_1–L_{26}) corresponds to the language identifiers defined in Table 5, which lists their full names and associated datasets for clarity. Table 5 also lists the IDs of all curated datasets containing recordings in each language, guiding researchers to their relevant data resources. Figure 3 shows the speech duration per language across all 50 datasets.

Observations: Figure 3 shows the total language duration across datasets. Languages such as L_2, L_{10}, L_{14}, L_{16}, L_{19}, L_{21}, L_{24}, and L_{25} have duration of more than 1000 hours. However, several languages such as L_3, L_5, L_{11}, L_{12}, L_{13}, L_{15}, L_{22}, and L_{23} urgently need more data. Furthermore, Table 5 reveals stark disparities in language coverage across tasks. Languages such as L_2, L_7, L_8, L_{10}, L_{14}, L_{16}, L_{17}, L_{18}, L_{19}, L_{20}, L_{21}, L_{24}, L_{25}, and L_{26} dominate, with at least 500 hours of speech for most tasks. These languages benefit from extensive corpora that yield hundreds of hours of usable audio. On the other hand, there is an urgent need of speech datasets in L_3, L_{11}, L_{12}, L_{13}, L_{15}, L_{22}, and L_{23} for all the tasks as current duration is less than 500 hours for each of these tasks. These patterns underscore that task-critical data is concentrated in a handful of languages, leaving many Indian languages under-served.

Table 5 reveals a highly imbalanced distribution of available speech data across tasks. Across tasks, T_3 (LID) and T_9 (GRE) frequently exceeds T_1 (MO-ASR/STT) and T_2 (ML-ASR/STT), indicating that multilingual pooling naturally provides LID and GRE ready data even when transcribed content is

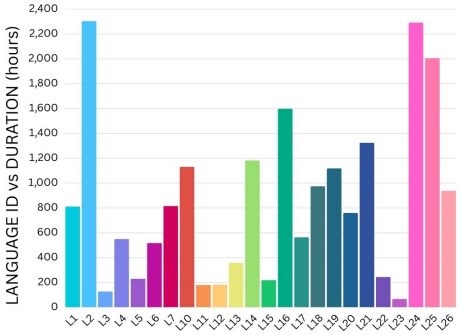

Figure 3: Total speech duration for each Indian language (L_1–L_{26}) across all 50 datasets. Language L_8 (Hindi) and L_9 (Indian English) have 3,981 and 16,154 hours of data and were excluded from the figure due to duration; they would have dominated the visualization and obscured relative differences among datasets. Languages like L_2, L_{24}, and L_{25} have the highest duration, whereas languages like L_{23}, L_3, L_{11}, and L_{12} are virtually absent.

sparse. In contrast, T_6 (SER) remains the least populated axis, while T_1, T_2, T_7 (MO-TTS), and T_8 (ML-TTS) collectively accumulate the majority of hours. Identical durations across T_1, T_2, T_7, and T_8 for several languages suggest shared corpora rather than distinct recordings.

Coverage under T_5 (ADD) is absent for L_3, L_{11}, L_{12}, L_{17}, L_{22}, and L_{23}; limited for L_4, L_5, L_9, L_{13}, and L_{20}; but comparatively higher for L_2, L_8, L_{10}, L_{14}, L_{16}, L_{19}, L_{21}, L_{24}, and L_{25}. Dravidian languages such as L_{10}, L_{14}, L_{24}, and L_{25} exhibit strong presence in T_3, T_5, and T_9, reflecting diverse annotations and sustained collection efforts relative to many Indo-Aryan counterparts.

Language-specific gaps remain evident. L_3 shows about 120 hours across tasks yet none in

Dataset	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
D_1	f_2	f_2, f_5	✓	✓	f_7	f_8	f_2	f_2, f_5	✓
D_2	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_3	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_4	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_5	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	f_9
D_6	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_7	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_8	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_9	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{10}	✓	✓	✓	f_6	f_7	f_8	✓	✓	f_9
D_{11}	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_{15}	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_{16}	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_{17}	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{18}	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_{19}	✓	f_5	✓	✓	f_7	f_8	✓	f_5	f_9
D_{21}	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_{22}	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_{23}	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{24}	✓	f_5	✓	✓	f_7	f_8	✓	f_5	✓
D_{25}	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{26}	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_{27}	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{28}	✓	f_5	✓	✓	f_7	f_8	✓	f_5	✓
D_{29}	✓	✓	✓	f_6	f_7	✓	✓	✓	✓
D_{30}	✓	f_5	✓	✓	f_7	f_8	✓	f_5	✓
D_{31}	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_{32}	f_2	f_2, f_5	✓	f_6	f_7	f_8	f_2	f_2, f_5	✓
D_{33}	✓	f_5	✓	f_6	f_7	f_8	✓	f_5	✓
D_{34}	✓	✓	✓	-	✓	f_8	✓	✓	✓
D_{35}	✓	✓	✓	✓	f_7	f_8	✓	✓	✓
D_{36}	✓	✓	✓	f_6	f_7	f_8	✓	✓	✓
D_{37}	f_2	f_2	✓	✓	f_7	f_8	f_2	f_2	✓
D_{38}	✓	f_5	✓	✓	f_7	✓	✓	f_5	✓
D_{39}	f_2	f_2	✓	f_6	✓	f_8	f_2	f_2	✓
D_{40}	✓	f_5	✓	✓	f_7	f_8	✓	f_5	✓
D_{41}	✓	f_5	✓	✓	f_7	✓	✓	f_5	✓
D_{42}	✓	f_5	✓	✓	f_7	✓	✓	f_5	✓
D_{43}	f_2	f_2, f_5	✓	✓	f_7	✓	f_2	f_2, f_5	✓
D_{44}	✓	f_5	✓	f_6	f_7	✓	✓	f_5	✓
D_{45}	f_2	f_2	✓	f_6	✓	f_8	f_2	f_2	✓
D_{46}	f_2	f_2	✓	✓	f_7	f_8	f_2	f_2	✓
D_{47}	f_2	f_2	✓	✓	f_7	f_8	f_2	f_2	f_9
D_{48}	✓	f_5	✓	-	✓	f_8	✓	f_5	✓
D_{49}	✓	f_5	✓	-	✓	f_8	✓	f_5	✓
D_{50}	✓	✓	✓	✓	f_7	f_8	✓	✓	f_9

Table 4: Dataset–task readiness summary. A ✓ denotes Task-Ready (all required features present); missing required features are listed as f_i, f_j. Tasks: T_1=MO-ASR/STT, T_2=ML-ASR/STT, T_3=LID, T_4=SV/SID, T_5=ADD, T_6=SER, T_7=MO-TTS, T_8=ML-TTS, T_9=GRE. *Notes:* Datasets D_{13}, D_{14}, and D_{20} are excluded because they are music-oriented rather than speech. Dataset D_{12} is a challenge release with extensive hours but insufficient public metadata for profiling, so it is omitted here. Although dataset D_{34}, D_{48}, and D_{49} includes f_6, which helps map a real speaker's voice to its synthetic counterpart, the dataset is better suited for anti-spoofing tasks than for speaker verification, where real and synthetic voices are matched.

T_4 (SV/SID), T_5 and T_6. A recurring observation is the dominance of shared datasets, particularly D_{15}, D_{18}, D_{39}, and D_{46}, which populate multiple cells in the matrix and span L_1, L_4, L_5, L_8, L_{13}, L_{15}, L_{16}, L_{18}, L_{19}, L_{24}, and L_{25}, across T_1, T_3, T_7, and T_8, demonstrating their multitask importance within the current resource landscape.

Language	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Datasets
L_1: Assamese	537	492	806	465	264	17	537	492	806	$D_{i\in\{2,3,8,15,18,29,35,39,46\}}$
L_2: Bengali	1990	1015	2297	635	1418	28	1990	1015	2280	$D_{i\in\{3–6,8,15–18,29,34–36,39,41–44\}}$ $D_{i\in\{49,50\}}$
L_3: Bhojpuri	120	120	120	0	0	0	120	120	120	$D_{i\in\{9\}}$
L_4: Bodo	483	483	550	487	63	0	483	483	550	$D_{i\in\{15,18,35,39,46\}}$
L_5: Dogri	200	200	231	204	26	0	200	200	231	$D_{i\in\{15,18,35,39,46\}}$
L_6: Garhwali	140	140	515	25	371	0	140	140	515	$D_{i\in\{9,35,39,46\}}$
L_7: Gujarati	810	810	810	409	197	0	810	810	799	$D_{i\in\{3,8,15,17,18,22,23,27,34,50\}}$
L_8: Hindi	3126	1243	3792	1138	1001	0	3126	1243	3689	$D_{i\in\{3,8,9,11,15–18,21,23–25\}}$ $D_{i\in\{34–36,39,45,46,50\}}$
L_9: Indian English	15940	32	15964	184	27	0	15940	32	15964	$D_{i\in\{1,25,26,31,40,46,48\}}$
L_{10}: Kannada	550	550	1124	195	837	0	550	550	1121	$D_{i\in\{3,8,15,17,18,25,27,34,35,39,46,50\}}$
L_{11}: Kashmiri	174	171	179	179	0	0	174	171	176	$D_{i\in\{18,19,35,46\}}$
L_{12}: Konkani	177	177	182	182	0	0	177	177	182	$D_{i\in\{15,18,35,46\}}$
L_{13}: Maithili	349	349	359	354	5	0	349	349	359	$D_{i\in\{15,18,35,39,46\}}$
L_{14}: Malayalam	681	679	1180	336	590	0	681	679	1174	$D_{i\in\{3,8,15–18,25,27,30,34–36,39,46,50\}}$
L_{15}: Manipuri	100	100	218	104	113	0	100	100	218	$D_{i\in\{15,18,35,39,46\}}$
L_{16}: Marathi	1103	994	1594	216	965	0	1103	994	1570	$D_{i\in\{3,7,8,15,17,18,23,27,34,35,39,46,50\}}$
L_{17}: Nepali	558	558	563	548	0	0	558	558	563	$D_{i\in\{3,4,6,8,15,18,35,46\}}$
L_{18}: Odia	650	650	973	208	410	0	650	650	973	$D_{i\in\{3,8,15,17,18,23,34–36,39,46\}}$
L_{19}: Punjabi	765	765	1115	228	728	0	765	765	1114	$D_{i\in\{3,8,15,17–18,34–35,39,46,50\}}$
L_{20}: Rajasthani	694	36	755	694	62	0	694	36	755	$D_{i\in\{15,28,39\}}$
L_{21}: Sanskrit	1321	1243	1325	133	999	0	1321	1243	1325	$D_{i\in\{15,17,18,33–35,46\}}$
L_{22}: Santali	240	240	245	245	0	0	240	240	245	$D_{i\in\{18,35,46\}}$
L_{23}: Sindhi	64	64	68	56	0	0	64	64	68	$D_{i\in\{8,15,18,35,46\}}$
L_{24}: Tamil	1630	1629	2178	762	1122	29	1630	1629	2156	$D_{i\in\{3,8,15,17,18,22,23,25,27,29\}}$ $D_{i\in\{34,35,38,39,45,46,50\}}$
L_{25}: Telugu	1389	1389	2005	711	966	0	1389	1389	2000	$D_{i\in\{3,8,15,17,18,22,23,25,27,34–36,39,46\}}$
L_{26}: Urdu	519	519	933	207	421	0	519	519	873	$D_{i\in\{3,8,10,17–18,32,34–35,39,46,50\}}$

Table 5: Language-wise Duration Distribution with Dataset Mapping (hours). The tasks with the highest and lowest durations for each language are highlighted in green and red, respectively.

5. Conclusion and Future Work

This paper introduced Task-Lens, a cross-task utility oriented lens for profiling speech datasets. Task-Lens involves four stages: dataset discovery, dataset filtering, feature extraction, and utility mapping. We conducted a comprehensive cross-task profiling of 50 Indian speech datasets spanning 26 languages and 91,257 hours of audio to assess their readiness for nine speech tasks based on the available metadata. Next, through Task-Lens outputs, we answered four vital research questions: (1) Which speech tasks does each dataset currently support? (2) What enhancements could improve cross-task applicability? (3) Which tasks lack sufficient data support? (4) Which Indian languages have adequate per-task coverage, and where do gaps persist? It turns out that datasets BhasaAnuvaad (Jain et al., 2024), South Asian Crowdsourced Speech (Kjartansson et al., 2018), IndicSUPERB (Javed et al., 2023b), IndicVoices-R (Sankar et al., 2024), MS Indic Speech (Microsoft, 2024), Nexdata AI 759h (Nexdata AI), NPTEL2020 (AI4Bharat, 2020), Rasa (Srinivasa Varadhan et al., 2024), IndicSynth (Sharma et al., 2025), NIRANTAR (Kumar et al., 2025b), EmoTa: A Tamil Emotional Speech Dataset (Thevakumar et al., 2025), IndicFake (Ranjan et al., 2025), BanglaSER (Das et al., 2022), SUBESCO (Sultana et al., 2021), BanSpEmo (Sultana et al., 2025), SEA_Spoof (Wu et al., 2025), IndieFake (Kumar et al., 2025a), Bangla Speech Corpus (Ahmed et al., 2020), and Indic-TEDST (Sethiya et al., 2024) lead the rankings. In contrast, many datasets need richer metadata diversity and greater scale. Similarly, language identification and gender recognition are moderately resourced. However, speaker verification / identification, audio deepfake detection, and emotion recognition remain critically underserved. From a language perspective, Bhojpuri, Dogri, Kashmiri, Konkani, Maithili, Manipuri, Santali, and Sindhi urgently need more data, as existing resources contain under 400 hours of speech.

This work opens several avenues for future research. By surfacing cross-task utility and highlighting dataset gaps, Task-Lens can give researchers immediate guidance on which dataset to use for their research problem, enabling them to focus on model innovation rather than time-consuming dataset curation. Furthermore, Task-Lens addresses inefficient use of existing resources and limited awareness of task appropriate datasets in

low resource languages. Generic AI tools and raw web searches list datasets based on their original intent. However, they rarely provide reliable cross-task profiles for efficient use. Additionally, by highlighting under-resourced languages and tasks, this work supports researchers interested in creating datasets for under-served areas. With over 7,000 languages spoken worldwide, extending Task-Lens can help advance inclusive speech processing research across diverse languages and tasks.

6. Limitations

Despite the breadth and utility of Task-Lens, we acknowledge certain limitations:

1. **Language and Task Coverage:** Our exploration is limited to nine core speech tasks, 26 languages and 50 Indian speech datasets, excluding less accessible datasets and emerging tasks (e.g., code-switching ASR). However, we plan to integrate additional community-contributed resources to broaden language and task support. Furthermore, all per-dataset feature values will be released upon publication to facilitate Task-Lens extension.

2. **Quality-aware utility estimation:** The current readiness check treats feature presence as binary. Incorporating signal quality, annotation quality, and basic sensitivity analyses can calibrate utility scores as well and strengthen their connection to downstream performance.

Despite these limitations, the NLP community will gain a practical, transparent foundation for efficient dataset selection, a clear roadmap for extending exploration to new languages, tasks, and standards, and practical insights into data resource gaps that can drive more resource contributions in the future[2].

7. Ethical Considerations

We encourage adoption of Task-Lens to support systematic cross-task profiling across diverse languages and tasks. Most datasets in our survey carry Creative Commons Attribution (CC BY 4.0) licenses; a smaller subset uses commercial terms or research-/academic-use-only terms (e.g., ELRA research, LDC-IL), alongside a few custom or paid licenses. All of these permit non-commercial cross-task research. One dataset (Wu et al., 2025) is released under CC BY-NC-ND 4.0, which restricts derivative uses; we nevertheless include it due to its uniquely valuable synthetic speech data for Southeast Asian languages. We will release verified license details for all included datasets upon publication to promote clarity and responsible reuse. Researchers should confirm each dataset's license and README, respect usage restrictions, and ensure compliance with privacy, consent, and data-protection norms.

[2]We used Grammarly and ChatGPT to assist with sentence construction and language refinement. While submitting our paper, we indicated it as a 20% use of generative AI.

8. Acknowledgments

This work is supported by the Infosys Centre for Artificial Intelligence (CAI) at IIIT-Delhi. We also thank the SBILab at IIIT-Delhi for their helpful discussions and support.

9. Bibliographical References

Mohamed T. Agro, Anuja Kulkarni, Karim Kadaoui, Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review. *arXiv preprint arXiv:2507.07741*.

Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2021. A toolbox for construction and analysis of speech datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Aarti Bakshi and Sunil Kumar Kopparapu. 2018. Spoken indian language identification: a review of features and databases. *Sādhanā*, 43:53.

Laurent Besacier, Etienne Barnard, Alex Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Carlos Busso, Mehmet Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Jianan Chen, Chenhui Chu, Sheng Li, and Tatsuya Kawahara. 2024. Data selection using spoken language identification for low-resource and zero-resource speech recognition. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Tokyo, Japan. APSIPA.

Somnath Dey, Md. Sahidullah, and Goutam Saha. 2023. An overview of indian spoken language

recognition from machine learning perspective. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):Article 128.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Cini Kurian. 2015. A review on speech corpus development for automatic speech recognition in indian languages. *International Journal of Advanced Networking and Applications*, 6(6):2556–2558.

Retno Larasati. 2025. Inclusivity of ai speech in healthcare: A decade look back.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaş, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. Dataperf: benchmarks for data-centric ai development. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

S. T. A. Monisha and S. Sultana. 2022. A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages. *Advances in Multimedia*, page 9602429.

Muhammad Bilal Mustafa, Muhamad Asyraf Yusoof, Hayder Kareem Khalaf, Ahmad A. R. M. Abushariah, Mohammad Lejla Mohd Kiah, Hoi Ngan Ting, and Saravanan Muthaiyah. 2022. Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, 12(19):9541.

Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjorn Hrobjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. 2021. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372:n160.

Nivedita Palia, P. Ahmed, Amita Dev, and Shyam Sunder Agrawal. 2013. Hindi speech corpora: A review. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Shakti Prasad Panda, Amiya Kumar Nayak, and Saroj Kumar Rai. 2020. A survey on speech synthesis techniques in indian languages. *Multimedia Systems*, 26(4):453–478.

Harshalata Petkar. 2017. Review of development of speech corpora and speech recognition research in hindi. *International Journal of Engineering Research and Applications*, 7(7):12–19.

Sampoorna Poria and Xiaolei Huang. 2025. Bhaasha, bhasa, zaban: A survey for low-resourced languages in south asia – current stage and challenges.

Divya Sharma and Arun Balaji Buduru. 2022. FAt-Net: Cost-effective approach towards mitigating the linguistic bias in speaker verification systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1247–1258, Seattle, United States. Association for Computational Linguistics.

Ivan Shilin, Liubov Kovriguina, Dmitry Mouromtsev, Gerhard Wohlgenannt, and Roman Ivanitskiy. 2018. A method for dataset creation for dialogue state classification in voice control systems for the internet of things. In *Proceedings of R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PR_LEAL 2017)*, volume 2233 of *CEUR Workshop Proceedings*, pages 96–106, Aachen, Germany. CEUR-WS.org.

Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh, and Vishal B. Waghmare. 2012. Indian language speech database: A review. *International Journal of Computer Applications*, 47(5):17–21.

Amitoj Singh, Virender Kadyan, Munish Kumar, and Nancy Bassan. 2020. Asroil: A comprehensive survey for automatic speech recognition of indian languages. *Artificial Intelligence Review*, 53(5):3673–3704.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Anand H. Unnibhavi and D. S. Jangamshetti. 2016. A survey of speech recognition on south indian languages. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1122–1126.

Vijeta Verma, Tomesh Verma, and Vinita Sahu. 2018. A survey based study of indian language speech database for speaker recognition. *International Journal of Engineering Research & Technology (IJERT)*, 3(20). ISNCESR–2015 Conference Proceedings.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2802–2812, New York, NY, USA. Association for Computing Machinery.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.

10. Language Resource References

Abraham, Basil and Goel, Danish and Siddarth, Divya and Bali, Kalika and Chopra, Manu and Choudhury, Monojit and Joshi, Pratik and Jyoti, Preethi and Sitaram, Sunayana and Seshadri, Vivek. 2020. *Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers*. European Language Resources Association.

Adiga, Devaraja and Kumar, Rishabh and Krishna, Amrith and Jyothi, Preethi and Ramakrishnan, Ganesh and Goyal, Pawan. 2021. *Automatic Speech Recognition in Sanskrit: A New Speech Corpus and Modelling Insights*. Association for Computational Linguistics.

Ahamad, Afroz and Anand, Ankit and Bhargava, Pranesh. 2020. *AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition*. European Language Resources Association.

Shafayat Ahmed and Nafis Sadeq and Sudipta Saha Shubha and Md. Nahidul Islam and Muhammad Abdullah Adnan and Mohammad Zuberul Islam. 2020. *Preparation of Bangla Speech Corpus from Publicly Available Audio & Text*. European Language Resources Association (ELRA).

AI4Bharat. 2020. *NPTEL2020 Indian English Speech Dataset*.

Anish Bhanushali and Grant Bridgman and Deekshitha G and Prasanta Ghosh and Pratik Kumar and Saurabh Kumar and Adithya Raj Kolladath and Nithya Ravi and Aaditeshwar Seth and Ashish Seth and Abhayjeet Singh and Vrunda Sukhadia and Umesh S and Sathvik Udupa and Lodagala V. S. V. Durga Prasad. 2022. *Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi*.

Billah, M. M. and Sarker, L. and Akhand, M. 2023. *KBES: A dataset for realistic Bangla speech emotion recognition with intensity level*.

Swathanthra Malayalam Computing. 2020. *Malayalam Speech Corpus*.

Alexis Conneau and Mingda Ma and Sarthak Khanuja and Yossi Zhang and Vitaly Axelrod and Sravana Reddy Dalmia and Jason Riesa and Christian Rivera and Ankur Bapna. 2022. *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*.

Das, Rakesh Kumar and Islam, Nahidul and Ahmed, Md. Rayhan and Islam, Salekul and Shatabda, Swakkhar and Islam, A. K. M. Muzahidul. 2022. *BanglaSER: A speech emotion recognition dataset for the Bangla language*.

Amitoj Diwan and Ramesh Vaideeswaran and Shruti Shah and Aditya Singh and Shachi Raghavan and Shubham Khare and Vinay Unni and Shyam Vyas and Ashish Rajpuria and Chiranjeevi Yarra and Ankur Mittal and Preeti K. Ghosh

and Preethi Jyothi and Kalika Bali and Vivek Seshadri and Sreyan Sitaram and Saurabh Bharadwaj and Jaideep Nanavati and Raghava Nanavati and Kannan Sankaranarayanan. 2021. *MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages*.

ELRA. 2018. *GlobalPhone 2000 Speaker Package Corpus*. ISLRN: 331-592-378-424-7.

ELRA. 2024. *Urdu Speech Recognition Corpus (Desktop)*. ISLRN: 739-446-795-223-8.

He, Fei and Chu, Shan-Hui Cathy and Kjartansson, Oddur and Rivera, Clara and Katanova, Anna and Gutkin, Alexander and Demirsahin, Isin and Johny, Cibu and Jansche, Martin and Sarin, Supheakmungkol and Pipatsrisawat, Knot. 2020. *Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems*. European Language Resources Association.

IIT-M. 2021. *Hindi–Tamil–English ASR Challenge*.

IIT Madras. 2025. *IndicTTS: Text-to-Speech Corpus for Indian Languages*.

Jain, Sparsh and Sankar, Ashwin and Choudhary, Devilal and Suman, Dhairya and Narasimhan, Nikhil and Khan, Mohammed Safi Ur Rahman and Kunchukuttan, Anoop and Khapra, Mitesh M. and Dabre, Raj. 2024. *BhasaAnuvaad: A Speech Translation Dataset for 13 Indian Languages*.

Tahir Javed and Sakshi Joshi and Vignesh Nagarajan and Sai Sundaresan and Janki Nawale and Abhigyan Raman and Kaushal Bhogale and Pratyush Kumar and Mitesh M. Khapra. 2023a. *Svarah: Evaluating English ASR Systems on Indian Accents*.

Tahir Javed and Janki Nawale and Sakshi Joshi and Eldho George and Kaushal Bhogale and Deovrat Mehendale and Mitesh M. Khapra. 2024. *LAHAJA: A Robust Multi-accent Benchmark for Evaluating Hindi ASR Systems*.

Tanvir Javed and Kanishk Bhogale and Aditya Raman and Pratyush Kumar and Anoop Kunchukuttan and Mitesh M. Khapra. 2023b. *IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages*.

Kalluri, Shareef Babu and Vijayasenan, Deepu and Ganapathy, Sriram and M, Ragesh Rajan and Krishnan, Prashant. 2021. *NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling*.

Oddur Kjartansson and Supheakmungkol Sarin and Knot Pipatsrisawat and Martin Jansche and Linne Ha. 2018. *Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali*.

Konduri, Samhita and Pendyala, Kriti V. and Pendyala, Vishnu S. 2024. *KritiSamhita: A machine learning dataset of South Indian classical music audio clips with tonic classification*.

Abhay Kumar and Kunal Verma and Omkar More. 2025a. *IndieFake Dataset: A Benchmark Dataset for Audio Deepfake Detection*.

Kumar, Anubhav and Chaitanya, Vikas and Sarkar, Rohan and others. 2025b. *Nirantar: A dataset for non-stationary background noise resilient speech recognition in Indian languages*.

Microsoft. 2024. *Microsoft Speech Corpus (Indian languages)*. Version 1.0; Contains conversational and phrasal speech data in Telugu, Tamil and Gujarati. Data provided by Microsoft and SpeechOcean.com; research-only use.

Microsoft. 2024. *Rajasthani Hindi Speech Data*.

Jagabandhu Mishra and Mrinmoy Bhattacharjee and S. R. Mahadeva Prasanna. 2023. *I-MSV 2022: Indic-Multilingual and Multi-sensor Speaker Verification Challenge*.

Md Mahadi Hasan Nahid and Md Ashraful Islam and Md Saiful Islam. 2018. *Bengali Speech Recognition - Bangla Real Number Audio Dataset*.

Nexdata AI. *759 Hours – Hindi(India) Scripted Monologue Smartphone Speech Dataset*.

OpenSLR. *SLR122: Kashmiri Data Corpus*.

Pothula, Aishwarya and others. 2025. *Shrutilipi-Anuvaad: Benchmarks and Baselines for Automatic Speech Translation and Subtitling in Indian Languages*.

Prabhu, Abhishek and Sahu, Debashis and Shanmuganathan, Sivaji and Prabhu, Rajath Ashok and Krishnan, R. and Murthy, Hema A. and others. 2023. *SPIRE-SIES: Spontaneous Indian English Speech (SIES) Corpus*. ArXiv:2312.00698.

Rambabu, Banothu and Gangashetty, Suryakanth V. 2018. *Hindi–English Code-Mixed Speech Corpus for ASR*.

Rishabh Ranjan and Mayank Vatsa and Richa Singh. 2025. *IndicFake Meets SAFARI-LLM: Unifying Semantic and Acoustic Intelligence for Multilingual Deepfake Detection*.

Ashwin Sankar and Srija Anand and Praveen Srinivasa Varadhan and Sherry Thomas and Mehak Singal and Shridhar Kumar and Deovrat Mehendale and Aditi Krishana and Giri Raju and Mitesh Khapra. 2024. *IndicVoices-R: Unlocking a Massive Multilingual Multi-speaker Speech Corpus for Scaling Indian TTS*.

Nivedita Sethiya and Saanvi Nair and Chandresh Kumar Maurya. 2024. *Indic-TEDST: Datasets and Baselines for Low-Resource Speech to Text Translation*. ELRA.

Sharma, Divya V and Ekbote, Vijval and Gupta, Anubha. 2025. *IndicSynth: A Large-Scale Multilingual Synthetic Speech Dataset for Low-Resource Indian Languages*. Association for Computational Linguistics.

Singh, Yeshwant and Biswas, Anupam. 2021. *Indian Regional Music Dataset*. Zenodo.

Singh, Yeshwant and Waikhom, Lilapati and Meena, Vivek and Biswas, Anupam. 2022. *Indian Folk Music Dataset*. Zenodo.

Sodimana, Keshan and Silva, Pasindu and Sarin, Supheakmungkol and Kjartansson, Oddur and Jansche, Martin and Pipatsrisawat, Knot and Ha, Linne. 2018. *A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese*.

Praveen Srinivasa Varadhan and Ashwin Sankar and Giri Raju and Mitesh M Khapra. 2024. *Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings*.

Srivastava, Nimisha and Mukhopadhyay, Rudrabha and K R, Prajwal and Jawahar, C V. 2020. *IndicSpeech: Text-to-Speech Corpus for Indian Languages*. European Language Resources Association.

Sultana, Babe and Hussain, Md Gulzar and Rahman, Mahmuda. 2025. *BanSpEmo: a Bangla audio dataset for speech emotion recognition and its baseline evaluation*.

Sultana, Sadia and Rahman, M. Shahidur and Selim, M. Reza and Iqbal, M. Zafar. 2021. *SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla*.

Syeda Mustafiza Tamim and Prangshu Manjul and Stephen Fernandes and Nithin S. and Roopashri M. R. and Narayan Kumar Choudhary and Shailendra Mohan. 2025. *Assamese Text to Speech Corpus*.

Thevakumar, Jubeerathan and Thavarasa, Luxshan and Sivatheepan, Thanikan and Kugarajah, Sajeev and Thayasivam, Uthayasanker. 2025. *EmoTa: A Tamil Emotional Speech Dataset*. International Committee on Computational Linguistics.

Ujjwal and Himanshu Garg and Mayank Joshi. 2020. *GACMIS: Genre Automated Classification using Machine Learning of Indian Songs*.

Vuddagiri, Ravi Kumar and Gurugubelli, Krishna and Jain, Priyam and Vydana, Hari Krishna and Vuppala, Anil Kumar. 2018. *IIITH-ILSC Speech Database for Indian Language Identification*. Satellite workshop of Interspeech 2018.

Jinyang Wu and Nana Hou and Zihan Pan and Qiquan Zhang and Sailor Hardik Bhupendra and Soumik Mondal. 2025. *SEA-Spoof: Bridging The Gap in Multilingual Audio Deepfake Detection for South-East Asian*.