INDICVOICES: Towards building an Inclusive Multilingual Speech Dataset for Indian Languages

Tahir Javed[1,2] **Janki Atul Nawale**[1] **Eldho Ittan George**[1] **Sakshi Joshi**[1,2]
Kaushal Santosh Bhogale[1,2] **Deovrat Mehendale**[1] **Ishvinder Virender Sethi**[1]
Aparna Ananthanarayanan[1] **Hafsah Faquih**[1] **Pratiti Palit**[1] **Sneha Ravishankar**[1]
Saranya Sukumaran[1] **Tripura Panchagnula**[1] **Sunjay Murali**[1] **Kunal Sharad Gandhi**[1]
Ambujavalli R[1] **Manickam K M**[1] **C Venkata Vaijayanthi**[1]
Krishnan Srinivasa Raghavan Karunganni[1] **Pratyush Kumar**[2,3] **Mitesh M Khapra**[1,2]

[1]AI4Bharat [2]Indian Institute of Technology Madras [3]Sarvam AI
{tahir, miteshk}@cse.iitm.ac.in

Abstract

We present INDICVOICES, a dataset of natural and spontaneous speech containing a total of 7348 hours of read (9%), extempore (74%) and conversational (17%) audio from 16237 speakers covering 145 Indian districts and 22 languages. Of these 7348 hours, 1639 hours have already been transcribed, with a median of 73 hours per language. Through this paper, we share our journey of capturing the cultural, linguistic and demographic diversity of India to create a one-of-its-kind inclusive and representative dataset. More specifically, we share an open-source blueprint for data collection at scale comprising of standardised protocols, centralised tools, a repository of engaging questions, prompts and conversation scenarios spanning multiple domains and topics of interest, quality control mechanisms, comprehensive transcription guidelines and transcription tools. We hope that this open source blueprint[1] will serve as a comprehensive starter kit for data collection efforts in other multilingual regions of the world. Using INDICVOICES, we build IndicASR, the first ASR model to support all the 22 languages listed in the 8th schedule of the Constitution of India.

1 Introduction

Recent advancements in Automatic Speech Recognition (ASR) have achieved remarkable success in English (Radford et al., 2023; Baevski et al., 2020, 2022; Chen et al., 2022; Gulati et al., 2020), with single digit WERs on multiple benchmarks (Panayotov et al., 2015; Meyer et al., 2020; Ardila et al., 2020; Hernandez et al., 2018; Wang et al., 2021). However, despite several massively multilingual efforts such as Whisper (Radford et al., 2023), USM (Zhang et al., 2023) and MMS (Pratap et al., 2023), the progress on mid and low resource languages is not at par with English. The primary reason for this is the lack of *sufficient*, *diverse* and *high quality* training data in these languages. In this study, we address this problem and set out with an ambitious goal to collect spontaneous speech data for Indian languages while respecting the linguistic, cultural and demographic diversity of India.

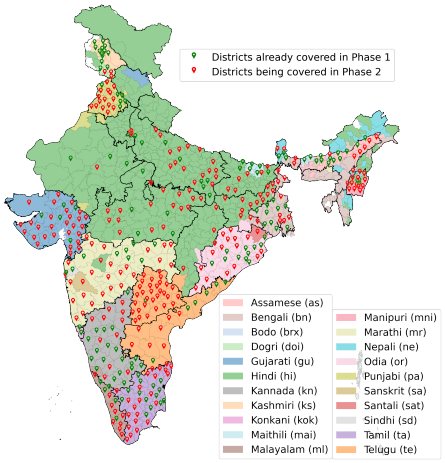

Figure 1: Primary regions of India where each of the 22 languages is spoken.

The scale of the problem is highlighted by the fact that there are 22 languages listed in the 8th schedule of Indian constitution, belonging to 4 different language families. These languages have a

[1]https://github.com/AI4Bharat/IndicVoices

[2]https://sites.google.com/view/gramvaaniasrchallenge/home/

[3]https://openslr.org/122/

[4]https://blog.smc.org.in/malayalam-speech-corpus/

[5]https://sites.google.com/view/indian-language-asrchallenge/home

Findings of the Association for Computational Linguistics: ACL 2024, pages 10740–10782

Dataset	#L	#Hours Ub	#Hours Lb	#Sp	#D	Type R	Type E	Type C	Channel WB	Channel NB
FLEURS (Conneau et al., 2022)	13	-	163	-	-	✓	✗	✗	✓	✗
MSR (Srivastava et al., 2018)	3	-	150	1286	1	✗	✗	✓	✗	✓
OpenSLR (Kjartansson et al., 2018)	3	-	618	1513	-	✓	✗	✗	✓	✗
CMS (He et al., 2020)	6	-	35	243	1	✓	✗	✗	✓	✗
MUCS (Diwan et al., 2021)	3	-	351	158	4	✓	✗	✗	✓	✗
Kathbath (Javed et al., 2023a)	12	-	1684	1218	3	✓	✗	✗	✓	✗
Shrutilipi (Bhogale et al., 2023a)	12	-	6457	-	-	✓	✗	✗	✓	✗
Graamvaani [2]	1	1000	108	-	-	✗	✗	✓	✗	✓
IISc-Mile (A et al., 2022b,a)	2	-	500	1446	-	✓	✗	✗	✓	✗
KDC [3]	1	-	1	43	-	✗	✗	✓	✗	✓
Vāksañcayah (Adiga et al., 2021)	1	-	78	27	8	✗	✗	✗	✓	✗
IIITH-ISD (Prahallad et al., 2012)	7	-	11	35	1	✓	✗	✗	✓	✗
IITB-MSC (Abraham et al., 2020)	1	-	109	36	1	✓	✗	✗	✓	✗
SMC-MSC [4]	1	-	2	75	4	✗	✗	✗	✗	✗
IITM [5]	3	-	690	-	-	✓	✗	✗	✓	✗
NPTEL (Bhogale et al., 2023b)	8	-	857	-	1	✓	✗	✗	✓	✗
IndicTTS (201, 2016)	13	-	225	25	4	✓	✗	✗	✓	✗
Svarah (Javed et al., 2023b)	1	-	10	117	37	✓	✓	✓	✓	✗
SPRING-INX (R et al., 2023)	10	-	2005	7609	16	✓	✓	✓	✓	✓
SPIRE-SIES (Singh et al., 2023)	1	171	23	1607	-	✗	✓	✗	✓	✗
IndicVoices	**22**	**7348**	**1639**	**16237**	**52**	✓	✓	✓	✓	✓

Table 1: Indian language datasets with number of languages (**#L**), hours of labeled(**Lb**) & unlabeled(**Ub**) data, number of speakers (**#Sp**), number of domains (**#D**), type of data (**R**ead/**E**xtempore/**C**onversation, wide band **(WB)**/narrow band**(NB)**).

collective speaker base of 1.2B, spread across 742 districts in India. Figure 1 shows the regions of India in which each of these languages is primarily spoken. The geographical spread of the 22 languages across India contributes to a rich diversity in culture, traditions, customs, beliefs, lifestyles, preferences and interests.

To collect data which is inclusive and representative of this diversity of India, we ensure a comprehensive representation across various demographics such as gender, age, educational background, and geographic location, with specific quotas for each category. Additionally, we aimed for diversity in the vocabulary and content, including a mix of read speech, voice commands, extempore discussions, and both wide and narrow-band recordings. We also aim to maintain a balanced representation of urban and rural speakers and recording devices and include a significant portion of data recorded in noisy environments representing everyday usage of ASR systems. This is indeed a one-of-its-kind effort ensuring inclusion (see Figure 2) while covering a large number of languages, districts, do-

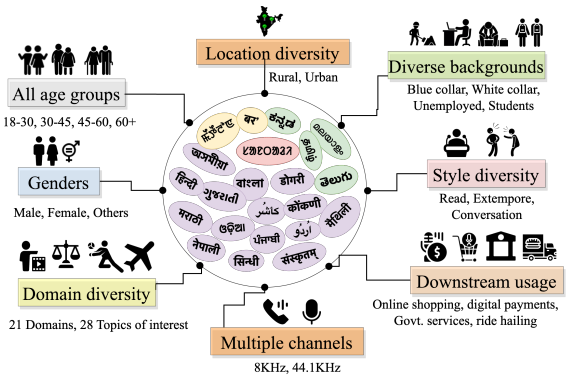

Figure 2: Demographic, geographical, domain, style and usage diversity wishlist for our data collection.

mains/topics with different recording channels and speech styles (see Table 1).

To achieve this, we created a clear framework for data collection which can be replicated across languages and diverse geographical locations. First, we created a frontend mobile application to enable data collection in a remote and distributed setup using a standardised procedure. Second, to elicit meaningful responses from participants capturing local contexts and culture, we created a repository of 2.5K engaging questions, 46.6K prompts, 1.1K (Dogri) to 4.1K (Hindi) role-play scenarios covering 21 domains and 28 topics of interest, anchored in practical, day-to-day usage scenarios. To facilitate data collection, we set up a countrywide network consisting of agencies, local universities, local NGOs, and social sector professionals who acted as regional influencers to engage a variety of participants. We also built an *in-house* quality control team who listened to every audio collected from the field and ensured adherence to a stringent acceptance criteria. Lastly, we built a team of transcribers consisting of makers, checkers and supercheckers and created an elaborate two-level transcription guideline addressing the unique transcription challenges in Indian languages.

IndicVoices is the result of this massive effort involving a total of 1893 personnel in a variety of roles such as language experts, local mobilisers, coordinators, quality control experts, transcribers, language leads and project managers. It contains a total of 7348 hours of read (9%), extempore (74%) and conversational (17%) audio data from 16237 speakers covering 145 districts and 22 languages, of which 1639 hours have already been transcribed (the rest is under progress). In addition to ASR, the collected data can be used for several other pur-

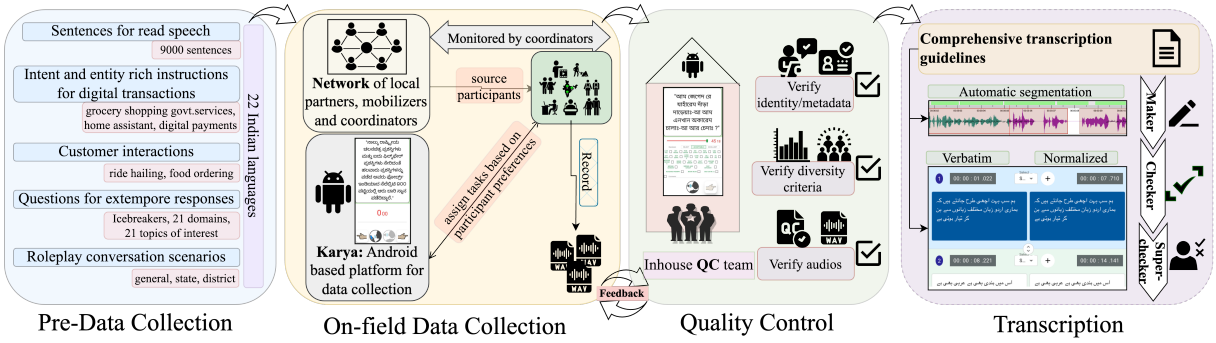

Figure 3: The four stages in the process used for collecting INDICVOICES. The collection and quality control is done using our in-house tool, an extension of Karya. The transcription is done using our in-house tool.

poses, such as speaker diarization, speaker identification, speaker verification, language identification, intent detection, entity extraction, query by example and audio denoising. We hope that all the data, tools, guidelines and other material created as a part of this work serve as an open source framework for data collection projects in other multilingual regions of the world. All artefacts developed in this work are open-source and can be accessed on GitHub[6].

In the remainder of this paper, we describe the four stages (Figure 3) in our process, *viz.*, (i) preparation before data collection (ii) on-field data collection (iii) quality control and (iv) transcription. We then present statistics about the data and train/evaluate existing models on INDICVOICES.

2 Pre-collection Data Preparation

Simple prompts often fail to engage participants effectively, necessitating the creation of structured sentences, questions, and scenarios for participants to read, respond to, or act out, as described below.

2.1 Sentences for read speech

For read speech, crucial for covering specific domain-related vocabulary, we curated approximately 9000 sentences from Wikipedia across 13 domains, translated into 22 languages by in-house translators. This ensures the inclusion of diverse and potentially challenging words not typically found in spontaneous speech. Participants are then asked to read these sentences verbatim, ensuring coverage of a wide range of topics and vocabulary.

2.2 Tasks involving voice assistants

Tasks involving voice assistants are essential for representing speech recognition applications in daily use. For such tasks, we devised specific instructions mirroring real-world interactions with digital services as described below. We then instructed participants to vocalize these instructions.

Everyday tasks: We utilized the MASSIVE (FitzGerald et al., 2022) dataset which contains interactions with home-assistants. We selected 9000 English sentences covering 60 intents and 18 domains, which were then translated into 22 Indian languages by in-house translators, with a focus on colloquial language and code-mixing to mirror typical usage patterns. Translators also localized content by substituting Western entities with Indian counterparts, enhancing cultural relevance.

Digital transactions: Representing interactions from *digital financial transactions*, *online grocery transactions*, and *government services* is crucial, given their prevalence and as they represent a broad range of similar use cases. Each of these three applications features distinct entities like bank names, product names, and government schemes. They also have unique characteristics, including numeric sequences, code-mixing, and application-specific terminologies. To accurately capture these interactions, we employed a standardized process (see Figure 4). We first identified common intents within each domain, such as *transferring money*, *buying groceries*, or *availing a government service*. Human annotators then crafted utterances in everyday language to express these intents, which were then abstracted into templates by substituting specific details with placeholders (e.g., numbers, product names, or service names). A comprehensive list of real-world entities was compiled, including numeric values, brand names, bank names, etc. Finally, we generated a variety of interactions for each language by substituting placeholders in the

[6] https://github.com/AI4Bharat/IndicVoices

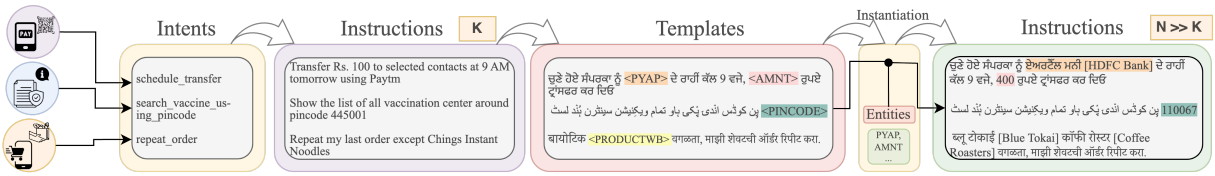

Figure 4: Standardised process for using a small number (k) of human-created instructions to create templates and then instantiate a large number of instructions (N) by replacing entities in the templates.

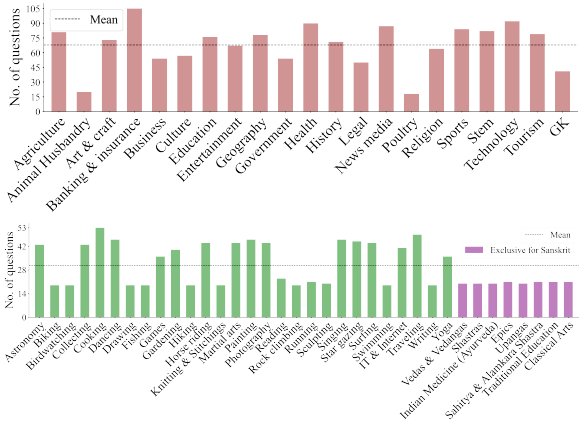

Figure 5: Number of engaging questions created for different domains (top) and topics of interest (bottom). These questions are contextualised and colloquially translated into 22 languages.

templates with these entities, ensuring a diverse and representative set. In each language, we created a total of 28K single turn instructions covering the three applications. These instructions would then be spoken out by participants with the flexibility to improvise as per their preferences.

2.3 Customer care interactions

The previously mentioned use cases involved scripted interactions designed for comprehensive entity, intent, and characteristic coverage, including code-mixing. Expanding this, we developed scenarios for participants to role-play interactions with customer care services of ride-hailing and food delivery apps, encouraging creative engagement rather than scripted responses. For instance, one scenario involves inquiring about car rental details for a family road trip. We crafted a total of 102 such scenarios covering the two services, translated into all 22 languages with cultural localization.

2.4 Extempore questions

Capturing natural speech through extempore conversations is challenging, as participants often struggle to engage with unfamiliar topics, leading to hesitant or superficial responses. To enhance engagement, we provided options from 21 domains and 28 topics, yet discovered that broad prompts like "talk about politics" were not effective. With the help of journalists, we created questions for every domain, which were accessible and relatable, including for rural audiences (see Figure 5). This rich set of questions sometimes resulted in repetitive answers; for instance, when asked to mention a local landmark in Madurai, nearly all responses spoke about the Meenakshi Amman Temple, neglecting other significant sites. To address this, we introduced hints aimed at directing participants towards less recalled yet meaningful topics, effectively broadening the scope of discussion. We also needed some special modifications (as highlighted in Figure 5) in the questionnaires for Sanskrit, which is not a language spoken by the masses.

2.5 Icebreaker questions

To create a comfortable dialogue environment, we developed three sets of easy-to-answer icebreaker questions, totalling 50-70 per set, designed to progressively warm up participants. These ranged from questions related to (i) daily lifestyle, (ii) household habits, and (iii) their mother tongue.

2.6 Questions about named entities

Downstream applications of speech recognition systems often necessitate the accurate identification of names, numbers, locations, and dates. Considering India's vast geographical and cultural diversity, which influences the variation in such entities across regions, we incorporated tasks asking participants to list entities like people names, Indian and international cities, dates, and numbers.

2.7 Role-play scenarios

Apart from the above monologues, it is also important to capture natural dialogues between two participants. However, initiating spontaneous conver-

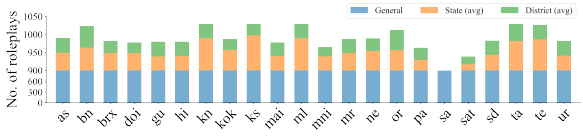

Figure 6: No. of general, state-specific and district specific roleplay conversational scenarios per language.

sations between strangers often led to basic, unengaging exchanges. To address this, we introduced structured roleplay prompts with specific roles and scenarios to enrich interactions. These prompts are divided into three categories: general, state-specific, and district-specific scenarios. General scenarios involve everyday interactions applicable across various regions, like *bargaining in a shop*. State-specific scenarios delve into conversations unique to particular states, reflecting their cultural or social nuances, such as a *discussion between a Kashmiri artisan and a local about handcrafted items*. District-specific roleplays focus on even more localized contexts, such as, *a conversation in Palakkad about the native varieties of rice*. Providing participants with such context-rich scenarios led to more natural and authentic conversations.

Please refer to Table 9 in Appendix B for examples of sentences, prompts, questions and role-play scenarios described in this section.

3 On-field Data Collection

We now describe the process of collecting data.

3.1 Countrywide Network

Collecting data from multiple districts across India would require assembling a network of district-specific influencers and mobilizers capable of recruiting participants across varied demographics. In addition, we would need local coordinators who can be trained in our data collection methodology to facilitate smooth data collection on-site. Initial pilots revealed that sourcing influencers, coordinators and participants for widely spoken languages in urbanized regions is relatively easier. However, for languages such as Dogri, Manipuri, Santali, etc., which are spoken in more remote areas, there are significant challenges due to the lack of local infrastructure and expertise. To address this, we partnered with a mix of organizations, including (i) data collection agencies (ii) foundations working for preservation of languages, (iii) social sector professionals with fieldwork experience, and (iv) universities with linguistics departments to devise localized solutions that are tailored to the unique conditions of different parts of the country.

3.2 Platform for data collection

To manage data collection in a remote set-up distributed across the country, we adopted Microsoft's open-source crowdsourcing platform, Karya (Chopra et al., 2019). Karya is an Android application which enables users to undertake micro-tasks, such as reading sentences, answering questions, or performing scenarios, with the flexibility to review and resubmit responses. We made a few changes to the base version of the app to (i) display micro-tasks grouped into categories defined in Section 2.1 to 2.6 (ii) display the total recording duration and (iii) display hints for questions to avoid repetitive answers (See section 2.4). For two-participant conversations, we employed a cloud telephony service to facilitate and record calls, ensuring data collection on a stereo 8kHz channel.

Our backend infrastructure mirrors Karya's, utilizing a box server architecture. Specifically, our setup includes 22 box servers (small dedicated virtual machines, one for each language) that communicate with the Karya application on participants' phones to provide prompts and save responses on the cloud. These boxes are connected to a master server used for uploading data. The master server also hosts Grafana for real-time visualization of collected and verified data. The boxes periodically sync with the master server to back up database states, send updates, and receive new data for collection or verification.

3.3 Mobilisation and Training

In each district, we enlist the support of 1-2 mobilizers through local partnerships, leveraging their influence to raise awareness and communicate the objectives and compensations involved. Additionally, we employ 5-7 coordinators per district, often young graduates fluent in the local language and skilled in interpersonal communication, to facilitate a comfortable environment for participants and ensure high-quality responses. A deliberate effort is made to maintain gender balance among coordinators to accommodate female participants effectively. These coordinators receive comprehensive training on using the Karya app, understanding the micro-tasks, and adhering to the quality control criteria, enabling them to efficiently manage participant interactions and secure valuable data.

3.4 Collection Procedure

Adapting to regional preferences, we offered two primary methods for participant engagement: assembling at a centrally located facility within the district or conducting sessions at participants' homes, with the choice largely dictated by the participants' convenience and travel limitations.

Onboarding participants: The onboarding process begins with local mobilizers identifying potential participants, followed by coordinators making contact to explain the data collection's objectives, expected duration, compensation details, and obtaining consent. Each participant completes a registration form that captures essential demographic information such as age, gender, language proficiency, and areas of interest. This detailed profiling ensures that the tasks assigned via the Karya app are well-aligned with the participants' backgrounds and preferences. For example, a participant interested in politics would receive prompts related to it, while someone interested in technology might be asked about the latest tech trends.

Assigning tasks: Once onboarded, participants are assisted in installing the Karya app on their personal devices, a choice that allows the collection of data across a diverse range of device types and microphone qualities. Each participant is assigned several micro-tasks involving each of the following activities: (i) answering basic icebreaker questions (ii) reading sentences (unless they can't read) (iii) vocalising instructions for everyday tasks and online transactions for digital payments, grocery shopping and government services (iv) interacting with customer care enacting a given scenario (v) answering extempore on questions from the chosen topics and domains of interest (vi) reviewing specific products purchased in recent past and (vii) engaging in three role-play conversations with fellow participants, one each from general, state-specific and district-specific. These diverse set of tasks give us ample opportunity to capture local culture as well as data for practical usage from each participant.

Recording data: Participants independently complete each micro-task, with coordinators present to ensure adherence to instructions and to assess audio quality. This includes verifying the audibility of responses despite potential background noise, relevance to the given prompts, and avoiding unnecessary repetition or deviation. Participants re-record a micro-task in case of quality issues.

Logging out: Coordinators aim for at least 20 minutes of quality audio per participant. Once the data is uploaded, participants are instructed to uninstall the app. The entire process typically spans 1-4 hours based on the participant's engagement level, with compensation aligned with local wage standards for a half-day's effort.

4 Quality Control

We set up an in-house quality control team comprising 3-5 experts per language for verifying the meta-data, diversity criteria and content.

Verifying meta-data: In the early stages of data collection, discrepancies between participants' voices and their registered age or gender information were noted, sometimes due to intentional misinformation or errors in form completion. Additionally, inconsistencies in voice across a participant's submissions suggested data contributions from multiple individuals under a single ID or instances of a single individual posing as multiple participants. To address these authenticity concerns, we introduced a video recording micro-task for participants to verbally confirm their details, allowing our quality control (QC) team to verify age, gender, and voice consistency visually and auditorily. Discrepancies led to data rejection. We added privacy safeguards ensuring videos were not downloadable, accessible only to female members of the QC team and promptly deleted post-verification. Recognizing cultural sensitivities, especially among female participants hesitant to record videos, we facilitated alternative live verification through a WhatsApp call with a female QC member.

Verifying diversity criteria: The QC team was responsible for checking that we meet all the diver-

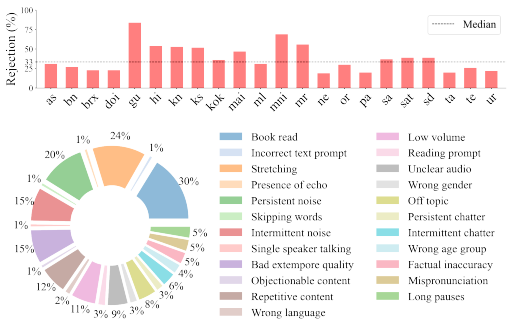

Figure 7: **Top:** Percentage of data that was rejected by our QC team for each language. The higher numbers are for languages where multiple pilots were done before we could find the right local partners. **Bottom:** Distribution of errors across the 23 categories.

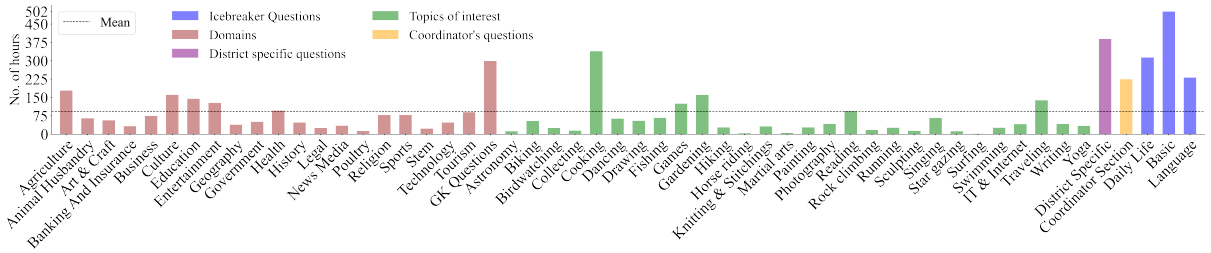

Figure 8: No. of hours of audio data collected from each of the domains and topics of interest summed up across languages. All domains have a good minimum representation, ensuring high diversity in the collected data.

sity criteria outlined in Figure 2. In case of a gap, the local partners were asked to source participants from specific underrepresented categories. The QC team also ensured that there was enough diversity in the domains and topics across participants. To ensure a broad range of topics, the QC team addressed the tendency of participants to choose familiar subjects like "Sports" and "Entertainment" by advising local partners to encourage discussions on less represented domains and topics. Lastly, the QC team would also monitor that easy conversation scenarios such as "teacher-student" or "customer-shopkeeper" were not repeated by disabling such options in the registration form. Figure 8 shows the diversity in content across topics and domains summed up across the 22 languages.

Verifying content: During initial data collection, our QC team noted issues with audio quality that were not caught by the first level of quality control conducted by coordinators. Problems included audio files with low volume and others with significant background noise, which compromised intelligibility. It was hypothesized that these issues slipped through due to coordinators' unfamiliarity with the context, allowing them to understand the content despite the poor audio quality. To address this, the QC team developed a detailed list of 23 error categories as shown in Figure 7 (also see Table 12 in Appendix E for a detailed description of the errors). Each audio file was assessed and categorized as Excellent, Acceptable (minor errors but transcribable by humans), or NotAcceptable (errors too severe for transcription). For audio files falling into the last two categories, the QC team identified and marked each specific error present in the file, ensuring a record of *all* issues detected.

5 Transcription

We now discuss the process of transcription, focusing on the guidelines and the workflow used.

Transcription guidelines: In Indian languages, there is often a significant divergence between spoken and written forms due to colloquialisms, accent variations, mispronunciations, and contractions in rapid everyday speech (see Appendix H. This issue necessitated a two-level transcription approach. In Level 1, we perform verbatim transcription, capturing spoken language as it is, despite deviations from standard written forms. In level 2, we standardized transcription, ensuring that all transcriptions conform to textbook representations, regardless of spoken variations, thereby improving usability in downstream applications. Given the high correspondence between grapheme and phoneme sets for Indian languages, Level 1 transcription guidelines simply require the transcribers to listen to the sound and transcribe it using the corresponding grapheme. However, Level 2 guidelines required more careful consideration, for which we collaborated with linguists to develop detailed rules for standardization. For each of the 22 languages, we crafted guidelines incorporating examples from the collected audio, undergoing several rounds of review to achieve consensus and finalize the guidelines, as detailed in Appendix H. We release both Level 1 and Level 2 transcripts giving researchers the flexibility to use as per their needs.

Transcription workflow: To maintain transcription quality, we employed a three-tier review system: maker-checker-superchecker. This process involved an extensive team of remote and external transcribers alongside experienced in-house super-checkers. The workflow, supported by our in-house tool, facilitated task assignment and monitoring across all 22 languages. The initial transcriptions (Level 1) were handled by the broader maker-checker team, while the super-checkers, as specialists, conducted secondary reviews and managed the transition to standardized Level 2 transcriptions, ensuring high accuracy and consistency.

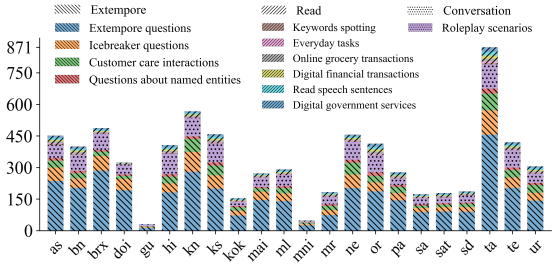

Figure 9: No. of hours of Read, Extempore and Conversation speech collected for each language along with splits for each sub-category.

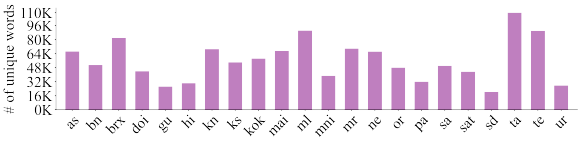

Figure 10: No. of unique tokens in the transcribed data across languages. The smaller bars correspond to languages for which transcribed data is very less.

6 IndicVoices

We collected a total of 7348 hours of data summed up across the 22 languages, covering 145 districts. Of these, 1639 hours of data have already been transcribed, with the rest in the pipeline. Figures 7 to 12 summarise different statistics of the collected data, which showcase its rich diversity.

Creating train-test splits: While creating train-test splits (Table 3), we ensured that the test set (i) does not contain any speakers which are already seen in the training data (ii) represents different age-groups, genders, education levels, professions, etc. (iii) represents speakers from every district and (iv) represents different downstream applications encompassing both narrow and wide band data. Creating such a balanced test set is challenging as it requires satisfying multiple constraints simultaneously. We relied on a sampling based method wherein we constructed multiple subsets of the data and then selected the one which satisfied

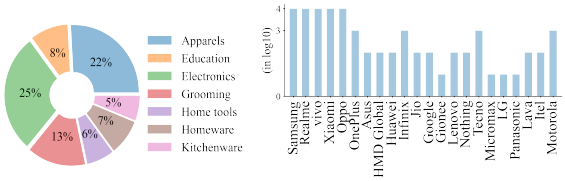

Figure 11: The LHS plot shows different categories in the product reviews spoken by participants. The RHS plot shows the number of recording devices from different brands that were used in data collection.

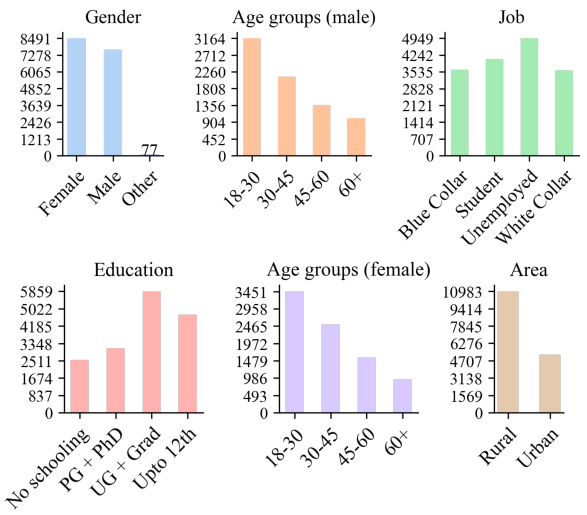

Figure 12: Demographic distribution of participants across categories summed up across all 22 languages.

all the constraints listed above, ensuring speaker exclusivity between train-test splits, and a balanced distribution of demographics, styles, applications and channels, allowing for fine-grained evaluation.

Evaluation: We start from a randomly initialized checkpoint and train a multilingual Conformer model, IndicASR, with 130 million parameters. Our model architecture follows the multisoftmax approach proposed by Tjandra et al. (2023) and is trained on IndicVoices using a hybrid CTC-RNNT objective with an alpha value of 0.3. We trained the model for 120K steps with a batch size of 512 audios per GPU. The max sequence length was set to 30 seconds. We used a learning rate of 1.0 and Noam as the learning rate scheduler. We evaluate this model on the IndicVoices test set and compare it with existing state of the art models, including Google USM (Zhang et al., 2023), Azure [7], Whisper (Radford et al., 2023), MMS (Pratap et al., 2023) and data2vecaqc (Baevski et al., 2022). We use Word error rate (WER) as the evaluation metric. As seen in Table 2, IndicASR is the first model to support all the 22 languages and it outperforms all existing models on supported languages.

Other tasks: Given the rich meta-data information, such as district id, speaker id, language id/proficiency, list of entities in a given interaction, etc., the collected data can be used for several tasks, including speaker diarization, speaker identification, speaker verification, language identification, intent detection, entity extraction, query

[7] https://azure.microsoft.com/en-in/products/ai-services/speech-to-text

Model	as	bn	brx	doi	gu	hi	kn	kok	ks	mai	ml	mni	mr	ne	or	pa	sa	sat	sd	ta	te	ur
USM	74.8	-	-	-	22.4	20.5	46.5	-	-	-	78.2	-	30.7	38.8	-	35.7	-	-	99.8	58.9	40.3	29.1
Azure	-	33.6	-	-	54.7	27.2	57.3	-	-	-	62.3	-	44.6	57.0	-	39.9	-	-	-	61.2	53.3	31.3
Whisper	127.2	93.2	-	-	60.5	34.0	96.5	-	-	-	148.6	-	95.2	100.8	-	89.1	99.0	-	92.6	78.4	151.9	29.7
Meta MMS	51.0	44.4	-	77.6	42.1	38.9	70.7	-	-	77.1	76.8	-	52.7	-	54.6	44.0	-	-	-	75.4	62.4	43.6
data2vec-aqc	-	60.5	-	-	38.5	27.8	59.2	-	-	-	77.2	-	33.6	-	-	44.7	-	-	-	56.8	-	36.5
IndicASR	20.4	15.9	25.9	33.3	21.4	15.0	30.3	31.4	39.3	33.6	40.5	26.1	18.2	16.4	23.4	12.9	22.8	35.4	29.6	31.2	26.8	14.4

Table 2: Performance of our IndicASR model compared with different models on the INDICVOICES benchmark.

Lang	Train		Test		Valid	Lang	Train		Test		Valid
	#sp	#h	#sp	#h			#sp	#h	#sp	#h	
as	684	100	163	5	1	mni	88	15	19	4	1
bn	590	100	125	5	1	mr	407	85	100	5	1
brx	694	100	164	5	1	ne	496	100	114	5	1
doi	321	65	84	5	1	or	391	75	92	5	1
gu	71	15	14	5	1	pa	274	71	61	5	0.4
hi	287	65	63	5	1	sa	176	45	33	5	1
kn	390	50	112	5	1	sat	237	86	53	5	1
ks	317	50	86	5	1	sd	108	10	30	5	1
kok	234	55	54	5	1	ta	863	100	195	5	1
mai	421	100	101	5	1	te	477	102	107	5	1
ml	332	60	67	5	1	ur	491	65	99	5	1

Table 3: Number of speakers (**#sp**) and hours (**#h**) in the train, validation and test and splits across languages.

by example, crosslingual transfer in multilingual ASR systems, multimodal speech-text embeddings. It can also be used for the evaluation of zero-shot transfer across districts with new accents. Given the episodic nature of our data collection, it can also be used for training and evaluating continual learning methods for ASR. This comprehensive dataset, thus, holds the potential to significantly advance a broad spectrum of speech and language processing tasks for multiple low-resource Indian languages.

7 Conclusion

We presented INDICVOICES, a comprehensive speech dataset adhering to an inclusive diversity wishlist with fair representation of demographics, domains, languages and applications. The materials developed as a part of this work will serve as a good starter kit for data collection for other languages. These materials include (i) Wikipedia sentences from multiple domains (ii) prompts for digital interactions, (iii) questions from different domains and topics of interest, (iv) conversational role-play scenarios (v) elaborate transcription guidelines (vi) android application for on-field data collection and verification and (vii) a web based platform for transcription workflow management. These can be adapted with minimal effort to bootstrap large scale data collection efforts for other languages.

8 Limitations

We list down the limitations of our work.

Number of districts: Due to budget constraints, for some languages such as Hindi and Urdu, which are spoken in multiple districts, we will have to skip about 40% of the districts (see Figure 1 for Hindi).

Number of speakers: Currently, we target 100-150 speakers per district. Ideally, we would have liked to target 4-5 times more speakers per district but this would have significant cost implications. We made a choice to collect minimum 20 minutes per speaker to ensure that there is enough diversity in the content as opposed to just collecting 4-5 minutes per speaker and targeting 4x-5x more speakers. The latter had two disadvantages: (i) the first 2-3 minutes of data would just end up being related to icebreaker questions, resulting in very little content diversity (ii) even if a speaker contributes just 4-5 minutes, we still have to pay half day's wage which would increase the total budget by 4x-5x.

Number of languages: In this work, we focus on 22 scheduled languages. Going forward we would want to extend this effort to many more languages.

Secondary districts: In this work, for a given language, we collect data only from districts where that language is identified as the primary language (i.e., spoken by majority of the population). Ideally, in subsequent phases, we would want to collect data from districts even if that language is spoken only by a small population in that district.

Topic Diversity: Despite our best efforts some topics such as *horse riding, sculpting*, etc. were underrepresented. This is understandable as very few participants could relate to such topics. Going forward, we will explicitly source participants who can speak on such underrepresented topics.

Two-party Conversations: Currently two-party conversations account for only 17% of our data. We would want this number to be at least 25%. The lower fraction is mainly due to the logistical difficulty in extracting meaningful conversations

from two unrelated participants.

Fine-grained evaluation: Given the limited space, we were not able to report a fine-grained evaluation across demographics and applications and just reported average WERs over the entire test set. We will fix this by conducting a shared task where we invite the community to participate and report fine-grained results for multiple models.

9 Ethics

The data collection process underwent rigorous ethical review and approval by the Institute Ethics Committee. Their recommendations, including the provision of all instructions in participants' native language, were duly implemented to ensure clarity and comprehension. Prior to recording, participants were fully informed about the purpose of data collection, the expected effort from their side, and the potential uses of their data. Their consent was obtained explicitly before proceeding. Participants were compensated appropriately, in line with the prevailing daily wages in their respective districts. This compensation was intended to recognize their time, effort, and contribution to the research endeavour. Careful attention was paid to participants' comfort and well-being throughout the data collection process. Refreshments such as tea, coffee, water, and biscuits were provided to ensure a hospitable environment conducive to participation. Participants were explicitly informed of their right to skip any task they felt uncomfortable with. The privacy and confidentiality of participants' personal identifiable information (PII) were rigorously safeguarded. No PII data will be shared externally, and measures were implemented to anonymize and protect sensitive information. Participants were given the option to have their videos deleted immediately after verification or to opt for a live call instead of recording a video for identity verification purposes. This choice was provided to accommodate individual preferences and ensure participant comfort and privacy. Lastly, all transcribers and other personnel employed in the project were also compensated appropriately according to prevailing salaries and wages. All tools will be released with an MIT license[8] and the dataset will be released with CC-BY-4.0 license[9], allowing commercial usage.

[8] https://opensource.org/licenses/MIT

[9] https://creativecommons.org/licenses/by/4.0/

10 Acknowledgments

Embarking on this mission was only possible due to the support of numerous organizations, individuals and members of the Indian language technology ecosystem. We would like to take a few sentences to thank all of them.

Sponsors/Donors: First and foremost, we thank the Ministry of Electronics and Information Technology (MeitY), Government of India, for setting up the ambitious Digital India Bhashini Mission with the goal of advancing Indian language technology. The human infrastructure comprising of a large team of translators, transcribers, reviewers and language experts who worked on this project were supported by the generous grant given by Digital India Bhashini Mission to IIT Madras to serve as the Data Management Unit for the mission.
We are indebted to Shri Nandan Nilekani and Shrimati Rohini Nilekani for believing in us and supporting our work through generous grants from EkStep Foundation and Nilekani Philanthropies. These grants were used for (i) supporting many of the students, research associates, and developers who worked on this project, (ii) fulfilling many of our compute needs, and (iii) recruiting project managers to oversee the massive pan-India data collection activity undertaken as a part of this work.
We thank Microsoft for their grant to support the creation of benchmarks for Indian languages. We thank the Centre for Development and Advancement of Computing, Pune (CDAC Pune) for access to its ParamSiddhi super-computer which was used for mining bitext pairs at scale.

IIT Madras: We thank Prof. V Kamakoti (Director, IIT Madras), Prof. Mahesh V Panchagnula (Dean, IIT Madras), Prof. Ravindra Gettu (Dean, IIT Madras) and Prof. Manu Santhanam (Dean, IIT Madras) for their constant encouragement and administrative support. In particular, we are thankful for the office space provided to AI4Bharat which houses some of our students, researchers, language experts and administrative team.

Indian language technology community: We extend our heartfelt gratitude to the expansive Indian language technology community,

comprising academia, startups, and the deep tech industry, both within India and across the globe. It is with immense gratitude that we acknowledge the incredible foundation laid by the giants of this community, whose pioneering work has paved the way for our endeavours. We are truly grateful for the knowledge, insights, and advancements that we have built upon, as we stand on the shoulders of these remarkable contributors. In particular, we thank Prof. Rajeev Sangal (Professor Emeritus, IIIT Hyderabad), Prof. Pushpak Bhattacharyya (IIT Bombay), Prof. Dipti Mishra (IIIT Hyderabad), Prof. Hema Murthy (IIT Madras), Prof. Umesh S (IIT Madras), Prof. Rajat Moona (IIT Gandhinagar), Prof. Ganesh Ramakrishnan (IIT Bombay), Partha Talukdar (Google Research India), Dr. Swaran Lata (MeitY), Dr. Sobha L (AU-KBC) and Dr. Ritesh Kumar (Dr. B.R. Ambedkar University) for their critical insights and constructive feedback in improving the translation guidelines used for creating the datasets released as a part of this work (we apologize if we have missed anyone).

Language Experts: We express our deepest gratitude to our exceptional and highly dedicated team of language experts, including translators, transcribers, reviewers, quality control experts, coordinators, linguists and data leads whose invaluable contributions have been instrumental in the creation of INDICVOICES. Their unwavering commitment to adhering to guidelines and their remarkable ability to work seamlessly as a cohesive unit, despite being geographically dispersed, is truly commendable. The quality and accuracy of the manual datasets developed as part of this endeavor owes much to their unwavering efforts. We extend our heartfelt thanks to every member of our remarkable language team for their outstanding dedication and invaluable contributions.

Administration Team: We are profoundly thankful to the remarkable individuals, Krishnan Karunganni S, Ambujavalli R, C V Vaijayanthi, Ravishankar Venkateswaran, for their exceptional dedication, patience, and extraordinary leadership in managing such an expansive team of data leads, language leads, coordinators and transcribers. Their unwavering commitment to orchestrating and guiding this diverse group of language experts is truly commendable. Through their exceptional organizational skills and expertise, they ensured seamless coordination and maintained the highest standards of quality throughout the data collection process. We also thank our support staff Shanthi S, Bhanumathy M, Bhavana R, Suganya Kumaresan, and Kalaivanan A, who helped with recruitment and procurement.

Development Team: We also thank our development team comprising of our in-house engineers, as well as, engineers from Tarento for building Shoonya which enabled all the manual translation work. In the absence of Shoonya, it would have been impossible to manage such a diverse team spread across the country working towards a common goal. We thank members of our development team for their patience in working with the language experts and building features that helped improve both the speed and quality of translation.

Partners: We would also like to thank our partners, viz. Kashmir University, Goa University, Aripana Foundation, Korou Foundation, Pragyam Foundation, Calcutta Foundation, Suchana, Desicrew, Shaip, Navana Tech, Samskrita Bharati, Nava Data, Rekhta Foundation and Dogri Sanstha who helped in the data collection and transcription process.

Last but not least, we thank the Almighty for giving us the courage to embark on this mission!

10.1 The Team Behind the Scenes

This work was possible because the efforts put in by all the remarkable individuals listed below.

Administrative Team

Name	*Designation*	*Affiliation*
Krishnan Karunganni S	Chief of Operations and Delivery	AI4Bharat
Ravishankar Venkateswaran	Delivery Head	AI4Bharat
Bhanumathy M	Recruitment	AI4Bharat
Suganya Kumaresan	Recruitment	AI4Bharat
Shanthi S	Operations	AI4Bharat
Mohanarangan A	Operations	AI4Bharat
Kalaivannan A	Operations	AI4Bharat
Saranya B	Operations	AI4Bharat

Project Managers

Name	*Involvement*	*Affiliation*
C Venkata Vaijayanthi	Data Collection	AI4Bharat
Ambujavalli R	Data Transcription	AI4Bharat
Manickam K M	Quality Control	AI4Bharat

Data Leads

Name	*Affiliation*
Hafsah Faquih	AI4Bharat
Pratiti Palit	AI4Bharat
Sneha Ravishankar	AI4Bharat
Saranya Sukumaran	AI4Bharat
Tripura Panchagnula	AI4Bharat
Sunjay Murali	AI4Bharat
Kunal Sharad Gandhi	AI4Bharat
Faizan Qadri	AI4Bharat
Bishnu Prasad Barman	AI4Bharat
Janki Atul Nawale	AI4Bharat

Karya Team

Name	*Designation*	*Affiliation*
Devbrat Anuragi	Project Associate	AI4Bharat
Eldho Ittan George	Project Associate	AI4Bharat
Sakshi Joshi	MS Student	AI4Bharat, IIT Madras
Tahir Javed	PhD Student	AI4Bharat, IIT Madras

Shoonya Team

Name	*Designation*	*Affiliation*
Aparna Ananthanarayanan	Manager	AI4Bharat
Ishvinder Virender Sethi	Manager	AI4Bharat
Kunal Tiwari	Backend Developer	AI4Bharat
Kartik Virendra Rajput	Full Stack Developer	AI4Bharat
Ayush Panwar	Full Stack Developer	AI4Bharat

Data Collection Partners

Language	*Name*	*Affiliation*
Assamese	Bikash Chandra	Pragyam Foundation
Bengali	Saumya Verma Mahima Verma	Calcutta Foundation Calcutta Foundation
Bodo	Bikash Chandra	Pragyam Foundation
Dogri	Dr. Preeti Dubey	Dogri Sanstha
Gujarati	Rajan KM	Desicrew
Hindi	Rajan KM	Desicrew
Kannada	Rajan KM	Desicrew
Kashmiri	Dr. Adil Amin Kak Nazima Mehdi	Kashmir University Kashmir University
Konkani	Pradnya Bhagat	Goa University
Maithili	Avinash Kumar	Aripana Foundation
	Dheeraj Kumar	Aripana Foundation
Malayalam	Rajan KM	Desicrew
Marathi	Rajan KM	Desicrew
Manipuri	Yasin	Korou Foundation
Nepali	Bikash Chandra	Pragyam Foundation
Odia	Saumya Verma Mahima Verma	Calcutta Foundation Calcutta Foundation
Punjabi	Dr. Sudarshan Iyengar Khushi Bhamra Nimisha Mahajan	IIT Ropar IIT Ropar IIT Ropar
Sanskrit	Shashanka Hatwar	Sanskrit Bharati
Santali	Sunil Murmu Kamala Murmu	Suchana Suchana
Sindhi	Nilima Motwani Nida Fazli	Rekhta Foundation Rekhta Foundation
Tamil	Rajan KM	Desicrew
Telugu	Rajan KM	Desicrew
Urdu	Nida Fazli	Rekhta Foundation

Translation Team

Language	*Name*	*Designation*
Assamese	Devanga Pallav Saikia	Language Lead, Senior Translator
	Bikash Chandra	Senior Project Manager
	Bishnu Prasad Barman	Translator
	Dimpi Sarma	Translator
	Bonya Baruah	Translator
	Bikash Chetia	Translator
	Kangkana Deka	Translator
	Lelina Barman	Translator
Bengali	Sounak Dutta	Language Lead, Senior Translator
	Shambhobi Ghosh	Senior Translator
	Srija Mukherjee	Translator
	Shreerupa Chattopadhyay	Translator
	Natasha Ahmed	Translator
	Kathakali Bhoumik Das	Translator
	Atrayee Dutta	Translator
Bodo	Prafulla Basumatry	Language Lead, Senior Translator
	Bihung Brahma	Senior Translator
	Bikash Chandra	Senior Project Manager
	Sidwma Brahma	Translator
	Sansuma Brahma	Translator
	Jeetumoni Basumatry	Translator
	Ria Borah Sonowal	Translator
Dogri	Preeti Dubey	Senior Project Manager
	Lalit Mangotra	Senior Translator
	Veena Gupta	Senior Translator
	Shashi Pathania	Senior Translator
	Anju Bala	Translator
	Monika chandel	Translator
	Kulbhushan Jasrotia	Translator
Gujarati	Pranav Pandya	Language Lead, Translator
	Jayesh Adhyaru	Translator
	Naresh Kapadia	Senior Translator
	Faiz Masi	Translator
	Jimal Patel	Translator
Hindi	Jaya Sarawati	Language Lead, Senior Translator
	Sufiya Pathan	Senior Translator
	Deepika Agarwal	Senior Translator
	Aakansha Dubey	Translator
	Neha Bhakal	Translator
	Ayesha Pereira	Translator
	Veda Bharti	Translator
Kannada	Anagha H. N.	Language Lead, Senior Translator
	Adithi Raveendranath	Translator
	Abhigna Joshi	Translator

Continued on next page

Translation Team (continued)

Language	*Name*	*Designation*
	Shivakumar R. M.	Translator
	Arun Kumar	Translator
	Goutham M	Translator
	T.R. Nagesh	Translator
Kashmiri	Vijay Wali	Senior Translator
	Shafi Shauq	Senior Translator
	Ambreen Farooq	Translator
	Meer Bismah	Translator
	Syed Samreen	Translator
	Sumaya Jehangir	Translator
	Nazima Mehdi	Senior Project Manager
	Ishfaq Nisar	Translator
Konkani	Pradeep Padgaonkar	Senior Translator
	Pradnya Bhagat	Senior Project Manager
	Sandesh Prabhudesai	Senior Translator
	Sharat Raikar	Senior Translator
	Anwesha Singbal	Translator
	Cia Fernandes	Translator
	Ashwini Kamat	Translator
Maithili	Sanjay Jha	Language Lead, Translator
	Avinash Kumar	Senior Project Manager
	Yogendra Pathak	Senior Translator
	Dr. Chandramani Jha	Senior Translator
	Vikas Vineet Jha	Translator
	Priyeshi Kumari	Translator
	Rahul Kumar Jha	Translator
	Vijay Deo Jha	Translator
	Manoj Kumar Pathak	Translator
	Tulika Swati	Translator
	Prashant Kumar Jha	Translator
	Nandan Kumar	Translator
	Kishore Keshav	Translator
	Sanjeev Kumar Jha	Translator
	Deepak Kumar	Translator
	Juli Jha	Translator
	Swati Jha	Translator
	Aditya Bhushan Mishra	Translator
Malayalam	Jebi Mariam Kurian	Language Lead, Translator
	Manoj Varma	Senior Translator
	C. V. Sudheendran	Senior Translator
	Jihad M.	Translator
	Jiza Mariam Kurian	Translator
	Ann Mary Thomas	Translator
	Srilekha Padmakuma Nambiar	Translator
Marathi	Kunal Gandhi	Language Lead,Translator

Continued on next page

Translation Team (continued)

Language	*Name*	*Designation*
	Paresh Prabhu	Senior Translator
	Vrinda Sarkar	Senior Translator
	Ranjana Pathak	Senior Translator
	Saee Kodolikar	Senior Translator
	Prasad Jog	Translator
	Shweta Deshmukh	Translator
	Bhushan Oke	Translator
	Neha Satish Bandekar	Translator
Manipuri	Reena Ashem	Language Lead, Senior Translator
	Yasin Khan	Senior Project Manager
	Chingtham Diana Devi	Senior Translator
	Diana Thingujam	Translator
	Jahir Hussain	Translator
	Sanju Pukhrambam	Translator
	Alfina Khaidem	Translator
	Kshetrimayum Momo	Translator
	Padmabati Achom	Translator
Nepali	Sunita Dahal	Language Lead, Senior Translator
	Bikash Chandra	Senior Project Manager
	Dhaka Ram Kafle	Senior Translator
	Lekhnath Chhetri	Senior Translator
	Tika Ram Rai	Senior Translator
	D. Ghimiray	Translator
	Dr Srijana Sharma	Translator
	Dr Khagen Sharma	Translator
Odia	Satyabrata Barik	Senior Translator
	Pramodini Pradhan	Senior Translator
	Sai Sudeep Das	Translator
	Abhishek Parija	Translator
	Suchishraba Sarangi	Language Lead
	Bhimasena Bhol	Translator
	Surendra Chandra Tripathy	Translator
Punjabi	Armin Virk	Language Lead, Translator
	Pallavi Kaushal	Translator
	Shallu Rani	Translator
	Parneet Kaur	Translator
Sanskrit	Harisha H. M.	Language Lead, Senior Translator
	Dr. Suresha	Senior Translator
	Suprith S.	Translator
	Sailaja Nittala	Translator
	Vasudev Aital	Translator
	Vivaswini	Translator
	Dr. Narayan Dutt Mishra	Senior Translator
Santali	Kamala Murmu	Senior Project Manager

Continued on next page

Translation Team (continued)

Language	*Name*	*Designation*
	Baren Kisku	Senior Translator
	Prasanta Kumar Hansda	Senior Translator
	Baburam Murmu	Senior Translator
	Sripati Tudu	Senior Translator
	Urmila Murmu	Translator
	Raju Mardi	Translator
	Churki Hansda	Translator
	Promila Hansda	Translator
	Sova Tudu	Translator
	Sanjiban Murmu	Translator
	Satya Hembram	Translator
	Guna Hembram	Translator
	Sagen Murmu	Translator
Sindhi	Armin Virk	Language Lead, Translator
	Dr. Nalini	Senior Translator
	Prakash Tejwani	Translator
	Bharati Chainani	Translator
	Karan Vanni	Translator
Tamil	Shakir Azeem	Language Lead, Senior Translator
	Leema Rajavarman	Senior Translator
	Shivapriya Murali	Translator
	Sharmila Grahadurai	Translator
	V Sayeelakshmi Rajaganapathy	Translator
Telugu	Shakir Azeem	Language Lead, Senior Translator
	Karuna Vempati	Senior Translator
	N. Sujatha	Senior Translator
	Srimoukthika	Translator
	Srilakshmi B.	Translator
Urdu	Dr. Irfan Ahmed	Senior Translator
	Nazima Mehdi	Senior Project Manager
	Aishwarya Diwakar	Translator
	Anwar Wajhiuddin	Translator
	Muhammad Anzar	Translator
	Hasan Akram	Translator
	Dr. Javaid Aziz Bhat	Translator
	Hafsah Faquih	Translator
	Habeebunnisa	Translator
	Mohammad Afaan	Translator
	Naziya Rasool	Translator

Transcription Team

Language	*Name*	*Designation*
Assamese	Daisy Devi	Superchecker
	Rima Saikia	Superchecker

Continued on next page

Transcription Team (continued)

Language	*Name*	*Designation*
	Angelina B. Dihingia	Superchecker
	Dewashree Ojah	Superchecker
	Harshita Talukdar	Superchecker
	Munmun Saikia	Superchecker
	Neelakshi Das	Superchecker
	Parag Kumar Deka	Superchecker
	Shilpi Gogoi	Superchecker
	Mukti Duwara	Superchecker
	Suni Jyoti Kalita	Superchecker
	Kangkana Deka	Superchecker
Bengali	Maamritha nandi	Superchecker
	Ayan Chatterjee	Superchecker
	Sayantani Banerjee	Superchecker
	Subhadeep Bhattacharjee	Superchecker
Bodo	Pronita Basumatary	Superchecker
	Jwngsar Brahma	Superchecker
	Bithika Khaklary	Superchecker
	Kenak Basumatary	Superchecker
	Pabita Basumatary	Superchecker
	Donald King Narzary	Superchecker
	Mainu Basumatary	Superchecker
	Mili Brahma	Superchecker
Dogri	Bansi Lal Sharma	Superchecker
	Rakesh Kumar	Superchecker
	Sunil Choudhary	Superchecker
	Atul Sharma	Transcriber
	Kamlesh Kumari	Transcriber
	Kumari Rita	Superchecker
	Rajesh Manhas	Superchecker
	Shagun Singh	Transcriber
	Skindya Devi	Transcriber
	Sunil Kumar	Transcriber
	Varinder Kumar	Transcriber
Gujarati	Charmi Soni	Superchecker
Hindi	Lalita Manoj Gehlot	Superchecker
	Sudeep Kumar Mishra	Superchecker
	Humera Begum	Superchecker
	Rakshit Ghai	Superchecker
Kannada	Anusha	Superchecker
	Sandhya Gopal Shetty	Superchecker
	Bharathi R Nayak	Superchecker
Kashmiri	Rinko Ji Koul	Superchecker
	Mohammad Asjad Khan	Superchecker
	Zargar Adil Ahmad	Superchecker

Continued on next page

Transcription Team (continued)

Language	*Name*	*Designation*
	Naira Farooq	Transcriber
	Uzma Nisar	Transcriber
	Sheeba Shafi	Transcriber
	Rehana Qasim Shah	Transcriber
	Rafia Nabi	Transcriber
Konkani	Anwesha Sigbal	Superchecker
	Sujatha kambli	Superchecker
	Kapila Surendra Desai	Superchecker
	Sumeda Ankit Sheldekar	Transcriber
	Ankita Anand Zamburlikar	Transcriber
	Cialini Fernandes	Transcriber
Maithili	Sanjeev Kumar Jha	Superchecker
	Juli Jha	Superchecker
	Sabita mishra	Superchecker
	Nimmi Kumari	Superchecker
	Pankaj Kumar Jha	Superchecker
	Baiju Kumar Jha	Transcriber
	Kanchan mishra	Transcriber
	Nandita Mishra	Transcriber
	Prabhat Kumar Jha	Transcriber
	Prabhu Nath Mishra	Transcriber
	Raghunath Mukhiya	Transcriber
Malayalam	Niveditha Varma	Superchecker
	Amritha Krishnan	Superchecker
	Ann Mary Thomas	Superchecker
	Rajitha K V	Superchecker
	Lipi Pushpakaran	Superchecker
Manipuri	Tongbram Jimi Singh	Transcription Reviewer
	Mousami Oinam	Transcriber
	Anil Chingakham	Transcriber
Marathi	Suma G	Superchecker
	Rashmi Arun Sathe	Superchecker
	Shraddha P. Prabhu	Superchecker
	Shweta Deshmukh	Superchecker
Nepali	Suraj Sharma	Superchecker
	Padam Parajuli	Superchecker
	Ram Jiwan Rai	Superchecker
	Bhakti Rai	Superchecker
	Mausam Sharma	Superchecker
	Bikash Chetri	Transcriber
	Furtengi Sherpa	Transcriber
	Keshav Sapkota	Transcriber
	Menuka Bhujel	Transcriber
	Sharmila Sharma	Transcriber

Continued on next page

Transcription Team (continued)

Language	*Name*	*Designation*
Odia	Rupanjali Badajena	Superchecker
	Abhipsa Mohanti	Superchecker
	Surendra Chandra Tripathi	Superchecker
	Pragyansmita Sharma	Superchecker
	Nibedita Sahoo	Superchecker
Punjabi	Sandeep Kaur	Superchecker
	Gurjeet Kaur	Superchecker
	Jaspal Singh	Superchecker
	Sandeep Singh	Superchecker
Sanskrit	Shriganesh Devaru	Superchecker
	Nagaranjan V	Superchecker
	N Sridhar	Superchecker
Santali	Suku Murmu	Superchecker
	Sunil Murmu	Superchecker
	Prashantha	Superchecker
	Sagen	Superchecker
	Ladam Murmu	Superchecker
	Biplab Tudu	Transcriber
	Ganesh Saren	Transcriber
	Sadhuram Hembram	Transcriber
	Sandhya Mandi	Transcriber
Sindhi	Heeral Kavlani	Superchecker
	Jyoti Bhatia	Transcription Reviewer
Tamil	Sumithra	Superchecker
	Savitha V	Superchecker
	Krishna Arvind	Superchecker
	Madhu Vanisree	Superchecker
	Hemalatha Venkatesh	Superchecker
	Usha M K V	Superchecker
	Petchiammal	Superchecker
Telugu	Ranjith	Superchecker
	Prudhvi Sagar	Superchecker
	N Janaki	Superchecker
	Herugu Deepak Iyengar Amrutha Lakshmi	Superchecker
	Radhe Shyam Salopanthula	Superchecker
	Suguru Sri Harsha Rao	Superchecker
Urdu	Syed Raqib Siraj	Superchecker
	Shahnawaz Alam	Superchecker
	Saad Ahmad Mirani	Superchecker
	Faizur Rehman	Superchecker

Quality Control Team

Language	*Name*	*Designation*
Assamese	Tonmoyee Bhuyan	QC Analyst
	Anima Chetry	QC Analyst
Bengali	Sunanda Sarkar	QC Analyst
	Anamika Das	QC Analyst
	Aditi Ghosh	QC Analyst
Bodo	Ria Borah Sonwal	QC Analyst
	Bashiram Basumatary	QC Analyst
Dogri	Rahul Singh	QC Analyst
	Kuldeep Kumar	QC Analyst
Gujarati	Miraba Yogendrasinh Chudasama	QC Analyst
	Yutika Amitkumar Chauhan	QC Analyst
	Richa Thaker Bhatt	QC Analyst
Hindi	Afifa Anjum	QC Analyst
	Rimsha Jairajpuri	QC Analyst
Kannada	Asha N	QC Analyst
	Jayaseetha M N	QC Analyst
	Roopkamal S	QC Analyst
Kashmiri	Tahir Ahmad Sheikh	QC Analyst
	Fatima Ashraf	QC Analyst
Konkani	Varsha Vishnu Garad	QC Analyst
	Pooja C Tople	QC Analyst
Maithili	Chitralekha Anshu	QC Analyst
Malayalam	Haritha M S	QC Analyst
	Anjali P Nair	QC Analyst
	Hanan A Vaheed	QC Analyst
Manipuri	Khumanthem Yuremba Meitei	QC Analyst
Nepali	Anima Chetry	QC Analyst
Marathi	Prachi Dnyaneshwar Dharashivkar	QC Analyst
	Akshay Sarjerao Talekar	QC Analyst
	Manpritkaur Gurmitsingh Ragi	QC Analyst
	Nikita Parolekar	QC Analyst
	Vishwanath Arjun Bhandwale	QC Analyst
Odia	Mohammed Irfan	QC Analyst
	Aliva Pradhan	QC Analyst
	Pranita Das	QC Analyst
	Ulka Antariksha	QC Analyst
Punjabi	Amandeep Singh	QC Analyst
	Kanchan Bala	QC Analyst
Sanskrit	Sujatha Bala	QC Analyst

Continued on next page

Quality Control Team (continued)

Language	*Name*	*Designation*
Santali	Aditi Ghosh	QC Analyst
Tamil	Ramya	QC Analyst
	Muthathal Subramanian	QC Analyst
	Elumalai Ellappan	QC Analyst
	Suganthi V	QC Analyst
	Hemavardhini R	QC Analyst
	Yuvaraj R	QC Analyst
	Gnanasoundari A	QC Analyst
Telugu	Peddareddygari Praneeth Reddy	QC Analyst
	Vakkapati Monika Chandana	QC Analyst
	Purnag	QC Analyst
	Srilakshmi Garimella	QC Analyst
Urdu	Sabiya Jan	QC Analyst
	Imtiyaz Ahmad Dar	QC Analyst
	Tariq Ahmad Sheikh	QC Analyst

Field Coordinators

Language	*Name*	*Designation*
Assamese	Manas Pratim Kalita	Coordinator
	Srijana Ojha	Coordinator
	Juhi Bayan	Coordinator
	Madhushree Saud	Coordinator
	Abinash Bordoloi	Coordinator
	Pangkhi Das	Coordinator
	Mayuri Deka	Coordinator
	Tanmoy Jyoti Bhuyan	Coordinator
	Tonmoyee Bhuyan.	Coordinator
	Tarun Chetia	Coordinator
	Munmi Borpatra Gohain	Coordinator
	Niloy Pratim Kashyap	Coordinator
	Gitika Barman	Coordinator
	Rikshikha Kashyap	Coordinator
	Sadhana Devi	Coordinator
	Biswajyoti Deka	Coordinator
	Mahesh Sharma	Coordinator
	Tarun Chetia	Coordinator
	Bhabani Sarmah	Coordinator
	Jadab Kalita	Coordinator
	Jayashree Saikia	Coordinator
	Bhaskar Gogoi	Coordinator
	Pujashree Saikia	Coordinator
	Luckey Lahon	Coordinator
	Rupjyoti Narah	Coordinator
	Pankaj Medhi	Coordinator

Continued on next page

Field Coordinators (continued)

Language	*Name*	*Designation*
	Nilutpal Saikia	Coordinator
	Pranab Borah	Coordinator
	Rupam Hazarika	Coordinator
Bodo	Narbilash Brahma	Coordinator
	Didwm Basumatary	Coordinator
	Nisha Daimary	Coordinator
	Punpun Basumatary	Coordinator
	Dhanjita Swargiary	Coordinator
	Nisha Daimary	Coordinator
	Punpun Basumatary	Coordinator
	Dhanjita Swargiary	Coordinator
	Kriti Deepa Brahma	Coordinator
	Didwm Basumatary	Coordinator
	Eragdao Brahma	Coordinator
	Albina Islary	Coordinator
	Hungama Narzary	Coordinator
	Jasmine Hajoary	Coordinator
	Chinki Narzary	Coordinator
	Aloka Muchahary	Coordinator
Dogri	Navedika Mishra	Coordinator
	Rakesh Kumar	Coordinator
Hindi	Sanjay Kumar Paswan	Coordinator
	Ashok Kumar Jha	Coordinator
	Ajay Shankar Jha	Coordinator
	Anjali Jha	Coordinator
	Menka Minu	Coordinator
	Bhuneshwar Tiwari	Coordinator
	Shrishti Pathak	Coordinator
	Shailesh Kumar Shukla	Coordinator
	Shivani Singh	Coordinator
	Anjani	Coordinator
	Akanksha Yadav	Coordinator
	Pooja Mishra	Coordinator
	Ram Kumar Dubey	Coordinator
	Nishant Dubey	Coordinator
	Anjali Mishra	Coordinator
	Tanu Yadav	Coordinator
	Sushil Kumar Ray	Coordinator
	Sanjay Kumar Paswan	Coordinator
	Vijay Shankar Jha	Coordinator
	Pooja Mishra	Coordinator
	Atul Sharma	Coordinator
	Shatrudhan Gupta	Coordinator
	Vikas Gir	Coordinator
	Ankit Singh	Coordinator
	Vineeta Shukla	Coordinator

Continued on next page

Field Coordinators (continued)

Language	*Name*	*Designation*
	Rajiv Thakur	Coordinator
	Gautam Govind	Coordinator
	Nitish Kumar	Coordinator
Kashmiri	Faizan Qadri	Coordinator
	Aneesa Khan	Coordinator
	Rayees Ahmad Lone	Coordinator
	Zubair Ahmad Bhat	Coordinator
	Faisal Ahmad Malik	Coordinator
	Rizwan Uz Zaman Wan	Coordinator
	Idrees Rehman Mir	Coordinator
	Yasir Mustaq Bhat	Coordinator
	Mehjabeena Nisar	Coordinator
	Sakeena Mohi Ud Din	Coordinator
Konkani	Manthan H Dessai	Coordinator
	Snehal Devanand Prabhu	Coordinator
	Santoshi Mahendra Bakal	Coordinator
	Venkatesh Prabhu	Coordinator
Maithili	Rajiv Thakur	Coordinator
	Gautam Govind	Coordinator
	Nitish Kumar	Coordinator
	Sandeep Kumar Jha	Coordinator
	Saurabh Jha	Coordinator
	Geethanjali Mishra	Coordinator
	Vijay Shankar Jha	Coordinator
	Hirendra Kumar Jha	Coordinator
	Saket Kumar	Coordinator
	Shankara Nand Jha	Coordinator
	Gajendra Narayan Jhan	Coordinator
	Mukesh Kumar Mishra	Coordinator
	Neetu Kumari	Coordinator
	Devashish Jha	Coordinator
	Sushil Kumar Ray	Coordinator
	Ashish Kumar Jha	Coordinator
	Rakesh Kumar Ray	Coordinator
	Dhiraj Kumar Jha	Coordinator
	Premalata Kumari	Coordinator
	Raj Kumar Jha	Coordinator
	Raghunath Mukhiya	Coordinator
	Alia Ali	Coordinator
	Krishna Mohan Kumar	Coordinator
	Shital Kumari	Coordinator
	Sanjay Kumar	Coordinator
	S	Coordinator
	Narendra Kumar Mandal	Coordinator
	Ashwni Kumar	Coordinator
	Rajan K Pandey	Coordinator

Continued on next page

Field Coordinators (continued)

Language	*Name*	*Designation*
	Abha Kumari	Coordinator
	Kiran Kumari	Coordinator
	Mohammad Sirajuddin	Coordinator
	Naveen Kumar	Coordinator
	Ranjan Kumar	Coordinator
Nepali	Susmita Gurung	Coordinator
	Dinesh Kharga	Coordinator
	Pukar Rai	Coordinator
	Tulshi Rai	Coordinator
	Chanda Khawas	Coordinator
	Pranoy Rai	Coordinator
	Ritu Rai	Coordinator
	Ajit Biswa	Coordinator
	Nishal Sharma	Coordinator
	Manisha Subba	Coordinator
	Sarita Lama Rai	Coordinator
	Bishal Cheetri	Coordinator
	Ranjeeta Lama Chhetri	Coordinator
	Bivek Sarki	Coordinator
Sanskrit	Shashanka Hatwar T V	Coordinator
	Divya Pandey	Coordinator
	Haripriya R Kulkarni	Coordinator
	Vidyadhare	Coordinator
	Dr. Rukmangadha	Coordinator

We also acknowledge the efforts of transcribers, translators, administrative teams, field coordinators that are working with our external partners

References

2016. Resources for indian languages.

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022a. Knowledge-driven subword grammar modeling for automatic speech recognition in tamil and kannada.

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022b. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada.

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyothi, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pages 2819–2826.

Devaraja Adiga, Rishabh Kumar, Amrith Krishna, Preethi Jyothi, Ganesh Ramakrishnan, and Pawan Goyal. 2021. Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5039–5050, Online. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023a. 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.

Kaushal Santosh Bhogale, Sairam Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M. Khapra, and Pratyush Kumar. 2023b. Vistaar: Diverse benchmarks and training sets for indian language asr. *ArXiv*, abs/2305.15386.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Manu Chopra, Indrani Medhi Thies, Joyojeet Pal, Colin Scott, William Thies, and Vivek Seshadri. 2019. Exploring crowdsourced work in low-resource settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, Karthik Sankaranarayanan, Tejaswi Seeram, and Basil Abraham. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *Proceedings of Interspeech*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. 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, Marseille, France. European Language Resources Association.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018.

TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation, page 198–208. Springer International Publishing.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023a. Indicsuperb: a speech processing universal performance benchmark for indian languages. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023b. Svarah: Evaluating English ASR Systems on Indian Accents. In *Proc. INTERSPEECH 2023*, pages 5087–5091.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55, Gurugram, India.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association. [link].

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

Kishore Prahallad, Naresh Kumar Elluru, Venkatesh Keri, S. Rajendran, and Alan W. Black. 2012. The iiit-h indic speech databases. In *Interspeech*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1, 000+ languages. *CoRR*, abs/2305.13516.

Nithya R, Malavika S, Jordan F, Arjun Gangwar, Metilda N J, S Umesh, Rithik Sarab, Akhilesh Kumar Dubey, Govind Divakaran, Samudra Vijaya K, and Suryakanth V Gangashetty. 2023. Spring-inx: A multilingual indian language speech corpus by spring lab, iit madras.

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

Abhayjeet Singh, Charu Shah, Rajashri Varadaraj, Sonakshi Chauhan, and Prasanta Kumar Ghosh. 2023. Spire-sies: A spontaneous indian english speech corpus.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 11–14.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. 2023. Massively multilingual ASR on 70 languages: Tokenization, architecture, and generalization capabilities. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

A The Journey

INDICVOICES has been a life changing experience for all of us involved in the project. Below we share some of our experiences with the reader.

Humble Beginnings in the Holy City of Madurai. Our journey commenced in Madurai, Tamil Nadu, famous for the Meenakshi Amman Temple. Seeking blessings from the *Devi*, we commenced our journey with high hopes and a clear vision. However, this pilot quickly became a reality check, challenging our assumptions at every turn, especially regarding participant mobilization. Despite the dense population of India, finding willing participants became unexpectedly difficult. Trust was a major hurdle; many were hesitant to share personal information during registration, fearing potential fraud, especially when digital transactions were mentioned. This skepticism slowed down the mobilization process significantly, making it challenging to achieve the desired diversity in age and gender ratios.

Additionally, the time commitment required from participants—sometimes extending up to four hours to complete the recording process—added another layer of complexity. This duration, much longer than anticipated, tested the patience and commitment of our participants. Hesitancy in speaking freely was another obstacle; many participants showed reluctance in opening up, leading to numerous retakes to capture responses that were natural and usable. This reluctance often resulted in responses that lacked depth and spontaneity, necessitating multiple attempts to elicit more meaningful dialogue. The culmination of these challenges not only extended the duration of our pilot but also highlighted the importance of building trust and ensuring clarity in communication to facilitate smoother data collection processes in the diverse linguistic landscape of India.

The Forgotten Generation. Right from the beginning, we were very clear that we wanted sufficient participant from the senior citizen age group to capture their rich life experiences and tap into their repository of knowledge about Indian customs, traditions and beliefs. However, addressing the underrepresentation of the senior demographic emerged as a significant hurdle. The limited presence of individuals aged 60 and above necessitated a reevaluation of our outreach efforts. Engaging with old age homes, senior citizen clubs, and conducting home visits became essential strategies to include this vital segment of the population. This challenge highlighted the importance of inclusivity and the need for tailored approaches to ensure diverse demographic participation. Our concerted efforts to involve senior citizens bore fruit in an unexpectedly delightful way, particularly notable in the diverse regions of India. A striking example of this success was observed in Jammu, where the older population displayed remarkable enthusiasm towards contributing to the project. This enthusiasm stemmed not just from their fluency in their native language (Dogri, in this case) but also from a deep-seated urge to preserve and pass on their linguistic heritage. Contrary to the challenges faced with other age groups and languages, senior citizens in Jammu and beyond became invaluable participants. Their eagerness to contribute not only enriched our dataset with authentic, nuanced language use but also underscored the critical role of senior citizens in safeguarding linguistic diversity.

The Need for a Guiding Hand. In the initial pilots, our fears about 'What if people don't speak' were confirmed. We learnt the hard way that eliciting fluent speech with good quality content from individuals in interactions with strangers posed an intriguing challenge. Recognizing the pivotal role of participant's comfort in encouraging natural dialogue, we acknowledged the importance of assigning dedicated coordinators to accompany participants throughout the data collection process. Through targeted training, these coordinators were equipped with effective communication skills to engage participants authentically. Consequently, we established a procedural framework to ensure coordinators' proficiency in guiding participants through the data collection procedure. The coordinators also helped the participants in overcoming technical challenges in installing the app, getting accustomed to the "record-verify-submit" workflow on the app and clarifying the expectation with respect to each microtask.

Back to the drawing board. After these initial pilots, we took a pause to critically reassess our data collection strategy. Insights from our initial pilots revealed that participants often provided repetitive answers, and everyday conversations failed to yield a diverse vocabulary and had very little coverage of names, numbers, entities, and brand names critical for downstream Automatic Speech Recognition (ASR) applications. The pilots also offered us a glimpse into people's interactions with technology and their expectations from it, such as placing or-

ders or completing digital transactions. We also realised that simple prompts like "talk about politics" fell short, and sparking a meaningful dialogue between strangers over a phone call proved challenging, often resulting in mere exchanges of pleasantries without covering a wide array of topics.

Recognizing these gaps, we returned to the drawing board for an in-depth pre-collection phase. Our goal was to collect sentences with rich vocabulary, craft engaging questions spanning various domains, and create scenarios that simulate everyday digital interactions. We also refined our process to include tasks that would elicit responses rich in numbers, named entities, locations, etc. and add role-play scenarios with detailed narratives to encourage dynamic conversations between two parties.

Weather Plays Spoilsport. After refining our approach, we headed to the extreme north of India, to the beautiful land of Kashmir. Excited as we were, we encountered the formidable challenge of the region's harsh winter. Starting our pilot in mid-November in Srinagar, we were keenly aware of the narrow operational window before the onset of heavy snowfall, which renders data collection nearly impossible from December to February. The unique weather conditions and the limited daylight hours significantly restricted our daily operations, allowing us only a brief period between 10 a.m. and 5:30 p.m. for recordings. This time constraint meant that each coordinator could only manage sessions with a maximum of two participants per day, highlighting the need for a more adaptable approach to meet our productivity and deadline goals. As we moved forward, it became crucial to innovate our data collection methods, shifting towards conducting recordings within the warmth and accessibility of participants' homes in the subsequent districts. This was a good learning for us and an early realisation that given the diverse geographical landscape of India, we will have to be mindful of weather conditions, be it self-imposed afternoon curfews during summer in West Bengal, mobility issues during monsoon in Kerala and Goa, floods in Northeast India, harsh winters in the northern parts of Punjab, Delhi, Kashmir, Jammu and so on. On a lighter note, while working on INDICVOICES, we have become experts on weather conditions in different parts of the country.

A Journey to the Remotest Parts of India. After covering Kashmir (Kashmiri) and Jammu (Dogri), our journey took us to the north-eastern parts of India to conduct pilot studies covering Assamese, Bodo, Manipuri, and Nepali. Initially filled with enthusiasm, our venture into Assam's relatively tranquil settings soon led us into the more secluded and challenging terrains of Bodoland. Here, the sparse population and limited access to resources questioned the feasibility of our project, presenting a stark contrast to our prior experiences. Bodoland, with its serene yet isolated landscape of traditional tribal huts and quiet, dimly lit pathways, intimidating silence with no vehicular noise offered a unique set of challenges. The initial low turnout at our designated collection site in Kokrajhar prompted us to rethink our approach to engaging with remote tribal communities. Their concerns over privacy and the openness to share opinions reminded us of the delicate balance required to ensure inclusivity while being mindful of socio-political or geographical hurdles.

Our journey further led us to Kalimpong in West Bengal, a region where the linguistic landscape shifts dramatically to predominantly Nepali speakers, diverging significantly from the Bengali culture of the state. This diversity within a single state highlighted the intricate patchwork of India's linguistic heritage. In response, we tailored our data collection approach, creating district-specific hints to engage participants in a manner that resonated with their unique linguistic identity.

Following this, the endeavour in Imphal West, Manipur, introduced us to a different set of challenges marked by remoteness and the complexities of operating amidst constant disruptions. The initiation of the pilot in the compact confines of a local hotel was just the beginning of a journey punctuated by curfews, riots, and internet shutdowns, which significantly delayed our progress. The decision to romanize text data to include older participants unfamiliar with the local script and continuing annotation work offline during internet shutdowns were necessary to adapt to the ever-changing conditions in the state.

This expedition across the remote parts of India was not just a journey through diverse geographical landscapes but a deep dive into the heart of India's linguistic diversity. Each region, with its unique challenges, taught us the importance of resilience, adaptability, and the profound value of including voices from every region of the country, no matter how remote.

The Rural Urban Divide. Following this, we did some pilots in rural districts in West Bengal covering two languages, Bengali and Santali. This

illuminated the stark urban-rural divide, presenting unique challenges and learning opportunities at every turn. As we ventured into the tribal regions to engage with Santali-speaking communities, the serene yet complex rural setting offered a vivid contrast to the bustling urban environments we had previously navigated. Conducting recording sessions in the midst of forests, under the canopy of trees, or in the humble backyards of huts, we were confronted with the realities of rural life: limited internet connectivity, the scarcity of participants fitting specific demographic profiles, unpredictable weather, and the reticence of women to participate.

These experiences shed light on the significant divide between urban and rural contexts, especially in terms of technological accessibility and the relevance of certain questions and use cases. It became apparent that some of our initial questions, designed with an urban mindset, were not resonant with the daily experiences of rural participants. For instance, the concept of "hailing a cab" was alien to many, revealing a disconnect in the applicability of our queries. This insight prompted us to revisit our approach and make our questions more closely aligned with the realities of rural life. We modified scenarios to involve "arranging for transport for cattle or food grains," among other adjustments, ensuring our questions and use-cases were relevant and relatable to the lives of our rural participants.

This re-calibration was not merely about changing the wording of questions but about cultural and contextual sensitivity in linguistic data collection. This journey through the contrasting landscapes of India reinforced the notion that to ensure inclusivity, we need to constantly change our assumptions, adapt and improve our processes.

The Silence Amidst the Noise. While rural areas offered a raw and unfiltered glimpse into India's linguistic diversity, urban settings introduced a different set of challenges. In bustling metropolises like Mumbai, the search for tranquil venues for recording sessions became a Herculean task, particularly in densely populated areas. The urban clamour and the scarcity of quiet spaces escalated the costs and complexities of data collection, underscoring the logistical hurdles unique to urban centres. Moreover, the enthusiasm for participating in data collection efforts was noticeably muted among the urban populace, especially among professionals leading busy lives. This apathy necessitated innovative mobilization strategies to engage a demographic that seemed distant from the cause. Despite these obstacles, the endeavour to capture the linguistic essence of India's urban centres was as crucial as that of its rural counterparts. Overcoming this silence and capturing voices amidst the noise became an essential part of our mission.

India, a land of many festivals. Navigating the vibrant maze of India's festivals proved to be one of the most colourful challenges in our data collection journey. In a country where each region celebrates its own set of festivals with fervour and devotion, scheduling work around these celebrations was akin to finding a needle in a haystack of holidays. From Durga Puja in West Bengal and Odisha, Ramzaan in Kashmir, Bihu and Pongal across other parts, to the universally celebrated Diwali, our calendar was a mosaic of cultural festivities.

The complexity of scheduling was humorously encapsulated in a conversation with one of our partners, which turned into a comedic back-and-forth of date dodging.

```
Partner: We can't start in May as it
is too hot that time of the year!

We: Ok, let's start in June then.

Partner: No, that would be difficult due
to the monsoon season. Both June and July
would be washed out, quite literally!

We: That looks bad. Then we should
definitely start in August.

Partner: But then we would have Ganesh
Chaturthi which is a very important
festival here. No participants would
turn up during that time.

We: Phew, what about September?

Partner: Schools and colleges will have
exams so we won't be able to use them as
venues (other options would be expensive).

We (frustrated): Okay, then I guess after
that we would have to wait for Dusshera
(October), Diwali (November), Christmas
and New Year (December) to also pass by.

Partner (with a straight face): Yes, that
would be ideal!
```

This humorous exchange underscored a significant reality of executing a project of this scale in India. It taught us the importance of flexibility, patience, and the ability to laugh at the seemingly impossible task of scheduling around the endless cycle of festivals. In the end, these challenges just became a part of our journey, making every successfully covered district feel like a festival in its own right.

Everything that could go Wrong. Amidst all the celebrations, we soon realised that in a remote and distributed setup involving a large number of people, ensuring quality is a challenging task. Early in the pilot phases, we observed participants abruptly stopping mid-speech or drifting into unrelated conversations with coordinators without stopping the recording, leading to fragmented audio and content corruption. Despite comprehensive guidelines and thorough training, the human element introduced unpredictability in task execution, with participants sometimes simply reading the questions/prompts instead of answering/enacting them. It became very clear that we need to have an in-house QC team whose task would be to listen to every audio file collected on the ground and tag such errors. We iteratively refined our error categorization, adapting to new types of errors as they were discovered.

In their eagerness to assist participants, coordinators sometimes went above and beyond, inadvertently scripting entire responses or conversations. These were then merely read aloud by the participants, transforming what was supposed to be a spontaneous exchange into a rehearsed performance! This would diminish the authenticity of spontaneous speech. To combat these issues and preserve data integrity, we introduced error categories like 'Bad extemporaneous' and 'Book read,' so that such content could be tagged. Similarly, in telephonic conversations, a tendency emerged for one participant to dominate, leading to minimal contributions from the other party, a phenomenon tagged as 'SST' (Single Speaker Talking) by our QA team. Categories like 'Stretching,' 'Repeating Content,' and 'Long Pauses' were introduced to counter verbosity and repetition, ensuring the recordings were content-rich.

In several places, the authenticity of participants' identities emerged as a significant challenge. Concerns were raised when the voice of a participant didn't seem to align with their reported age and/or gender, leading to discoveries of intentional misinformation or unintentional errors in registration. Some instances revealed inconsistency in voices under the same participant ID, hinting at multiple individuals sharing a single ID, while others showed the same voice across different IDs, indicating individuals masquerading as multiple participants. To address these authenticity issues, we introduced a micro-task requiring participants to record a video stating basic information, allowing our QC team to verify age and gender visually. Disparities led to data rejection, while voice mismatches in audio samples triggered further scrutiny. Recognizing privacy concerns and cultural sensitivities, particularly among female participants reluctant to record videos, we offered alternatives like live verification through WhatsApp calls, conducted by female QC members, ensuring a respectful and secure verification process.

Data collection across diverse settings—outdoors, in public schools, small hotels, and participants' homes—brought the challenge of background noise interference. Distant ambient noises were less intrusive compared to the constant buzz of fans in closed spaces. Distinguishing between unavoidable natural background sounds and disruptive persistent noises was essential.

An unexpected challenge arose with the capture of highly expressive, albeit profane, reactions to daily frustrations, necessitating a balance between authenticity and appropriateness. This led to the creation of an 'Objectionable Content' category to carefully screen for hate speech or inappropriate content. Through this iterative process of reviewing audio files our QA team came up with 23 error categories which comprehensively captured everything that could go wrong!

The Subtle Art of Transcription. While we continued on our journey across the country, little did we know that our greatest challenge lay not in the fieldwork, but in the nuanced art of transcription. Initially perceived as a straightforward task of converting speech to text, the complexity of transcribing the diverse speech styles of thousands of individuals soon became apparent. The main issue was the different between colloquially spoken language and standardised language found in textbooks. The former contains words which may not have any spellings in standard textbooks or dictionaries but still cannot be ignored, simply because this is how people talk! This required us to choose between the pure phonetic representation of speech resulting in non-standard spellings on one hand, and pure textbooks representations

which deviated from what was being said on the other. To address this we ended up with a two-level transcription approach wherein the first level the transcribers were asked to transcribe verbatim without worrying about correctness of spellings (thus emphasising only on phonetic fidelity). In the second level the transcribers were asked to standardise the transcription to convert the phonetically correct representations to nearest standard spellings in textbooks. Thus the lazily spoken Hindi word "muje" would be transcribed verbatim as "muje" in Level 1 ensuring phonetic fidelity and then standardised to "mujhe" in Level 2 ensuring spelling accuracy. However, aligning all language experts and transcribers with this novel framework proved challenging. Many transcribers initially resisted typing verbatim spellings, feeling it betrayed the standard writing style. Through extensive discussions, iterative rounds of feedback, and a collaborative effort across language teams, we established a set of guidelines that balanced standardized systems with fidelity to the actual sound wave. This meticulous process underscored transcription not just as a task but as an art form, requiring a deep understanding of linguistic nuances, cultural context, and the delicate balance between preserving the integrity of spoken language and adhering to standard linguistic conventions.

Stories from the Heart of India. Despite the hurdles, the journey was largely filled with numerous positive experiences. The diversity we encountered in weather, languages, and cultures was overwhelming, yet it filled our hearts with an indescribable warmth and a profound appreciation for India's rich cultural tapestry. The love and hospitality offered by the people, their eagerness to share their stories, and their enthusiasm for preserving their linguistic heritage were truly heartening. Our journey underscored the power of language as a bridge to understanding people, their cultures, and the nation at a deeper level.

In a world where technology often seems to isolate us, our project brought people closer together, allowing them to connect and share their lives through their native languages. This technical endeavour became a conduit for genuine human connection, enabling people from various backgrounds to express their lives, traditions, and experiences. From a Tamil Nadu entrepreneur sharing her journey to success, to a Kashmiri woman finding solace in prayer during tumultuous times; from a young boy in Kashmir divulging his secret recipe, to a Manipuri girl aspiring for higher education amidst challenges, each story added a unique voice to the rich mosaic of IndicVoices.

The diversity in stories we collected — ranging from personal achievements and cultural narratives to expressions of socio-political concerns — highlighted the importance of the dataset we were building. Whether it was an old man reminiscing about life seventy years ago, a professor discussing the significance of a local folk culture, or a young Nepali girl playfully sharing tales of her dates, these narratives painted a vivid picture of the diverse life across India. Such stories not only enrich our understanding but also celebrate the myriad facets of Indian life, from its challenges to its triumphs.

Miles to go. Despite the vast disparities in lifestyle, language, and social circumstances across different regions, a common thread of respect for linguistic heritage and a passion for technology united participants from all walks of life. This shared enthusiasm underscores a collective commitment to preserving India's linguistic diversity, bridging the gap between tradition and modernity. IndicVoices, thus, stands as a testament to the enduring spirit of India and its people, weaving together the voices of its many inhabitants into a vibrant tapestry of stories that resonate with authenticity, diversity, and a profound sense of belonging. Yet, as extensive as our journey has been, it is but a chapter in a much larger story. With over 15,000 hours of recordings still ahead, our expedition through the heart of India's linguistic landscape is far from over. Like the timeless verse, "miles to go before I sleep" our path stretches onward, promising more voices to be heard, more stories to be shared, and an ever-deepening appreciation for the rich mosaic of Indian culture and language. This journey has only just begun, and the road ahead is filled with the promise of discovery, understanding, and the celebration of India's incredible diversity.

Figure 13: Karya's Home screen (left), micro-task screen (middle) and QC screen (right)

B Pre-data collection

Table 8 shows the number of sentences for read speech, number of instructions for different digital interactions, number of scenarios for customer care interactions, number of questions and role-play scenarios for two-party conversations. This required a team of linguists experienced in fieldwork for eliciting meaningful responses and translators for colloquially translating and contextualising the material to all the 22 languages. All of this material, released as a part of this work, can be adapted with minimal effort for data collection in other languages.

C On field Data Collection

Table 11 gives a summary of the micro-tasks to be completed by each participant. Note that participants who are not fluent in reading, can skip the read speech tasks. Such participants are also assisted by the coordinators in reading and understanding the questions, prompts, scenarios, etc. Figure 13 shows the screenshots of our Android application that was used for data collection and verification.

D Instructions for participants

The following instructions were given to the participants:

Aim of the Project: The aim of this project is to collect data for the development and evaluation of speech technology specifically tailored to your language. [The details of what is entailed by "speech technology" was colloquially explained by the coordinators to the participants using everyday applications.]

Purpose of Data Collection: Your data will be utilized for both commercial and non-commercial purposes to enhance speech technology for your language.

Amount of Time: The entire process is expected to take between 1 to 4 hours.

Compensation: You will receive compensation of INR X for your participation. The exact value of X will be communicated to you by the coordinator.

Consent: By participating, you agree to the terms and understand that your data will be used as mentioned above. Please proceed only if you consent to these terms and sign the consent form.

Registration: Your details will be collected during registration, ensuring your privacy is preserved. None of this information will be shared with third parties. You should also upload your signed consent form as a part of the registration process.

App Installation: Search for the "<Anonymous>" application on the Google Play Store using your Android smartphone and download the application.

Login: Use the access code assigned to you during registration to log in. Confirm this code with your coordinator before proceeding further.

Fetch Tasks: Press the "Submit Tasks/Get New Tasks" button to access the list of tasks assigned to you.

Recording: Click on each task and carefully read the prompt provided. If you find any task unclear, consult your coordinator for clarification. Take your time to formulate your response. If you are uncomfortable with a question, feel free to skip it. Take breaks as needed and complete the tasks at your own pace.

Submit a Task: After recording your response, click on "Stop". The application will then replay your response. If you and/or the coordinator are satisfied with the response, click on the "Next" arrow. Otherwise, click on "Record" to re-record your response.

Two-party conversations: Once you have finished the tasks on the App, the coordinator will pair you with a participant for conversation on a specific scenario. You can request to be paired with a participant of your choice and choose a role as well as scenario from a long list of roles scenarios that will be shared by the coordinator.

Logging Out: Once you have finished all tasks, return to the home screen and click on "Submit Tasks". Wait for a message confirming that all tasks have been submitted. You will then be prompted to record a video for identity verification purposes.

Type	as	bn	brx	doi	gu	hi	kn	kok	ks	mai	ml	mni	mr	ne	or	pa	sa	sat	sd	ta	te	ur
Sentences for read speech	9146	9200	9199	8976	9200	9040	9006	9119	9135	9196	9200	8444	9065	9200	9199	8982	9199	9152	5036	9200	7721	9068
Everyday tasks	9330	9359	9359	8000	8958	6363	9354	9093	5297	7962	9357	9356	9354	9350	9355	8171	9342	9346	7988	9359	9345	9302
Digital financial transactions	9448	9448	9448	9448	9448	9448	9448	9447	9448	9448	9448	9448	9448	9448	9448	9448	9448	9448	9444	9448	9448	9448
Online grocery transactions	9335	9335	9335	9335	9335	9335	9335	9334	9335	9335	9335	9335	9335	9334	9335	9335	9335	9335	9331	9335	9334	9335
Digital govt. services	9274	9274	9274	9274	9274	9274	9274	9273	9274	9274	9274	9274	9274	9274	9274	9274	9274	9274	9274	9274	9274	9272
Customer care interactions	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103	103
Extempore questions	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	2401	1216	2401	2401	2401	2401	2401
Icebreaker questions	152	152	152	152	152	152	152	152	152	152	152	152	152	152	152	152	145	152	152	152	152	152
Questions about named entities	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Role-play scenarios	1185	1911	201	196	887	3256	1355	448	423	439	662	199	1488	426	2700	329	000	245	425	1882	1804	2722
																			(+900 general roleplays)			

Table 8: Count of different kind of prompts and sentences, prepared as a part of our data collection for all the 22 languages.

Read Speech	**Sentences**
Wikipedia Sentences	*Source:* A Superintendent of Police is responsible for the police administration of each district. *Translation (mai):* एकटा पुलिस अधीक्षक प्रत्येक जिलाक पुलिस प्रशासन लेल उत्तरदायी होइत अछि। *Transliteration:* ikat pulis adhishak prateyek jela pulis prasasan lel uttardai hoita aich
Read Speech	**Commands**
Everyday tasks	*Source:* Can you change the light colors into dark in the house *Translation (hi):* क्या तुम घर में हल्के रंगों को गाढ़े रंगों में बदल सकते हो *Transliteration:* kya tum ghar mein halke rangon ko gaadhe rangon me badal sakte ho
Digital financial transactions	*Source:* Block the HDFC bank Holdings from my account. *Translation (kok):* म्हज्या खात्यांतलें एच.डी.एफ.सी बँक [HDFC bank] चे होल्डिंग्स ब्लॉक कर. *Transliteration:* mhajya khatyantle HDFC bank che holdings block kar
Online grocery transactions	*Source:* Stationary items delivered yesterday had expired. *Translation (ks):* سٹیشنٔری ہِندؠ چیٖز یِم راتھ وأتی تِم أسی ایٚکسپایر گٔمِتی *Transliteration:* Stationary hind cheez yim raath waeit tim aeis expire gemit
Digital government services	*Source:* I want EPFO passbook statement for account with UAN number 679696225561 *Translation (sd):* मूंखे UAN नंबर 679696225561 वारे खाते लाइ EPFO पासबुक स्टेटमेंट खपे। *Transliteration:* moonkhe UAN number 679696225561 vaare khaate laai EPFO passbook statement khape
Customer care	**Interactions**
Ride hailing	*Source:* Action: Book Ola/Uber auto (mention location) [Scenario: Book an auto/cab to go a specific location. Ask if the driver will be ariving soon.] *Translation (pa):* ਐਕਸ਼ਨ: ਬੁੱਕ ਓਲਾ/ਉਬੇਰ ਆਟੋ (ਸਥਾਨ ਦਾ ਜ਼ਿਕਰ ਕਰੋ) [ਦ੍ਰਿਸ਼: ਕਿਸੇ ਖਾਸ ਸਥਾਨ 'ਤੇ ਜਾਣ ਲਈ ਇੱਕ ਆਟੋ/ਕੈਬ ਬੁੱਕ ਕਰੋ। ਪੁੱਛੋ ਕਿ ਕੀ ਡਰਾਈਵਰ ਜਲਦੀ ਆ ਰਿਹਾ ਹੈ।] *Transliteration:* Action: book ola/ober auto (sthaan da zikar karo) [drish: kise khaas sthaan 'te jan layi ik auto/cab book karo. pucho ke ki driver jaldi aa riha hai]
Extempore	**Questions**
Domains and Topics	*Source:* What are some of your suggestions to improve healthcare in your area? What are the things that need to be changed? *Translation (ta):* உங்கள் பகுதியின் சுகாதாரத்தை மேம்படுத்த சில பரிந்துரைகளைச் சொல்லுங்கள்?என்னென்ன விஷயங்கள் மாற்றப்பட வேண்டும்? *Transliteration:* Ungal paguthiyin sugaathaaraththai membaduththa sila parinthuraigalaich sollungal? Ennenna vishayangal maatrappada vendum?
Icebreaker	*Source:* How often do you visit nearby towns/villages and for what purposes? *Translation (gu):* નજીકનાં નગર/ગામડાંની મુલાકાત તમે કેટલા સમયાંતરે લો છો અને કયા હેતુઓ માટે? *Transliteration:* NajiknaaN nagar/gaamDaaoni mulaakaat tame keTla samayaantare lo chho ane kayaa hetuo maaTe?
Questions about	**Named entities**
	Source: Can you name 5 states of India? *Translation (sat):* ᱵᱷᱟᱨᱚᱛ ᱨᱮᱭᱟᱜ ᱕ ᱜᱚᱴᱟᱝ ᱯᱚᱱᱚᱛ ᱨᱮᱭᱟᱜ ᱧᱩᱛᱩᱢ ᱠᱚᱢ ᱞᱟᱭ ᱫᱟᱲᱮᱭᱟᱜ-ᱟ ? *Transliteration:* Bharot reyag 5 gotang ponot reyag nutum kom lay dareyag-a ?
Roleplay	**Scenarios**
	Source: Rice seller- Buyer: Kaluniya and Badshavog rice prices are increasing *Translation (bn):* চাল বিক্রেতা | ক্রেতা | কালুনিয়া ও বাদশাভোগ চালের দাম বাড়ছে *Transliteration:* chaal bikretaa- kretaa: kaalunia o baadshaabhog chaaler daam baarche

Table 9: Examples of different sentences, prompts, questions and role-pay scenarios created as a part of our pre-collection preparation. These are localised and colloquially translated to all the 22 languages.

TV	animal	baby	baby_crying	tsk	barking	beep	bell	sniffle
child	child_crying	child_laughing	child_talking	dishes	child_yelling	children	children_talking	tones
click	clicking	clink	clinking	cough	child_whining	door	footsteps	gasp
horn	hum	inhaling	laughter	meow	motorcycle	music	nose_blowing	noise
phone_vibrating	popping	pounding	screeching	rattling	ringing	rustling	scratching	printer
smack	sneezing	sniffing	bird_squawk	snorting	squawking	squeak	stammers	static
throat_clearing	thumping	tone	children_yelling	trill	baby_talking	typewriter	ugh	uhh
wheezing	whispering	whistling	yawning	yelling	buzzer	clanking	phone_ringing	hmm
unintelligible	buzz	clanging	persistent-noise-start	tapping	singing	talking	umm	hiss
breathing	chiming	groan	persistent-noise-end	sigh	swallowing	uh-huh	siren	

Table 10: Different noise tags supported by Shoonya in transcription.

Task	Description
Sentences for read speech	4 sentences from Wikipedia that need to be read as it is.
Everyday tasks	4 commands used for interaction with in-home personal assistant.
Digital financial transactions	4 interactions that are typically encountered in digital transactions
Online grocery transactions	4 interactions that are typically encountered in such transactions
Digital government services	4 interactions that are typically encountered while interacting with such services
Keywords	10 keywords to be spoken as it is by the participant
Customer care interactions	2 scenarios to enact, one each from ride hailing and food delivery
Extempore questions	4 questions, two each from the participant's interested domain and topic of interest. The task requires participants to answer the questions in their natural way.
Icebreaker questions	3 questions, one from each tier, viz. warm-up, everyday-life and questions on participant's mother-tongue.
Product reviews	One positive review for a product that they have purchased recently and one negative review for a product that they have purchased recently.
Questions about named entities	7 questions asking about entites like names of countries, international cities etc.
Role-play scenarios	3 scenarios for conversations between two participants, one each from general, state-specific and district-specific. These are not done through the app but by connecting two participants on a telephone call through a bridge. Note that it is not necessary that the same person gets paired with the same partner for all the 3 conversations.

Table 11: Number of tasks assigned to each participant across each category and a brief description of the tasks

Error Type	Explanation
Low Volume	Audio not loud enough for clarity.
Intermittent Noise	Background sounds disrupting audio at intervals.
Persistent Noise	Constant background noise throughout the audio.
Intermittent Chatter	Background conversations at intervals.
Persistent Chatter	Continuous background conversations.
Unclear Audio	Poor audio clarity.
Off Topic	Content deviating from the prompt.
Repetitive Content	Unnecessary repetition in the audio.
Long Pauses	Unusually long silences in the audio.
Mispronunciations	Incorrect pronunciation of words.
Reading Prompt	Participant reading the prompt instead of responding.
Book Read	Monotonous, scripted reading detected.
Single Speaker Talking	Only one speaker present in single-speaker tasks.
Stretching	Unnatural elongation of words or sounds.
Objectionable Content	Inappropriate or offensive material in the audio.
Incorrect Text Prompt	Mismatch between text prompt and expected content.
Factual Inaccuracy	Inaccuracies in the information provided.
Skipping Words	Omission of crucial parts of the prompt.
Wrong Language	Incorrect language used for the task.
Presence of Echo	Echo affecting sound quality.
Bad Extempore Quality	Poor naturalness and fluency in extempore tasks.
Additional Comments	Other noteworthy observations not covered above.

Table 12: Error categories identified by the QC team

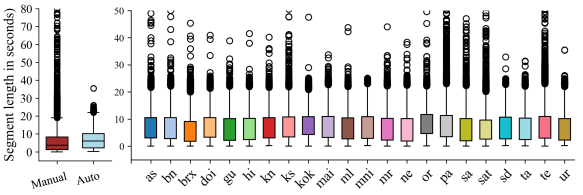

Figure 14: **Left:** Comparison of the distribution of segment lengths generated by manual segmentation and automated segmentation. On average, the manual segmentation process leads in very short (small mean) but also results some excessively large segments. **Right:** With the automated approach we are able to control the distribution of the length of audio segments ensuring a similar average length across languages and very few outliers. The outliers typically correspond to very fast spoken content where pauses are hard to identify.

This video will not be shared publicly and will be deleted once verified by our Quality Control team. If you are uncomfortable recording a video, inform your coordinator to arrange a WhatsApp call with a female member of our central team. This call will not be recorded, ensuring your privacy.

E Quality Control

Table 12 gives a brief description of the different error categories identified by the QC team. Table 10 refers to the different noise stage supported in Transcription

F Audio Segmentation Before Segmentation

Our data includes read, extempore, and conversational audio data, with read speech typically lasting 6 to 8 seconds, and extempore and conversational data extending up to several minutes. The extended duration necessitated segmentation due to the computational constraints of Automatic Speech Recognition (ASR) systems, optimized for 20 to 30-second audio contexts. Initially, we relied on annotators to listen to the entire audio and then delineate logical segments using an annotation tool. However, defining what constitutes a "logical segment" proved ambiguous, leading to considerable variability in the segmentation outcomes, with some annotators creating excessively short or long segments (as shown in Figure 14 (Left)).

To standardize segmentation and reduce subjectivity, we shifted to an automated approach using Silero Voice Activity Detection (VAD) (Team, 2021). This method determines voiced segment boundaries based on a specified minimum silence duration (S_{min}). The selection of S_{min} directly influences segment length, with shorter S_{min} values generating more segments by identifying every minor pause, and longer S_{min} values requiring more substantial pauses for segmentation. We followed an iterative approach, wherein we varied the value of S_{min} from 1000 ms to 40 ms. During each iteration, we retained segments which were less that 20 seconds and removed them from further processing. In subsequent iterations, longer audio files were further segmented by relaxing the minimum silence duration. Further, to prevent skew towards brief segments, we merged shorter segments, achieving an average segment length of 9 seconds. Figure 14 compares the distribution of the segments length across languages when using the automated approach.

G Creating train-test splits

To efficiently manage the complexity of satisfying multiple constraints, we create train-test splits using a sampling based method. More specifically, from the data collected for a given district, we draw a large number of samples that include all data from a random set of 20% of speakers from that district (referred to as test speakers). Each such sample may have a different set of speakers with its own demographic distribution. We then evaluate the demographic distribution of each sample, selecting the sample whose demographic distribution has the highest entropy and the lowest Kullback-Leibler (KL) divergence from the target demographic profile. We ensure that the data for none of the speakers in the chosen sample is present in the training data. Our process thus ensures a fair and robust benchmark ensuring speaker exclusivity between train-test splits, inclusion of all districts and a balanced demographic distribution, to the extent possible.

H Level 2 Transcription Guidelines for Hindi

To access L2 Transcription guidelines for all 22 Indian languages, please log in to the following link[10]

Rules for correcting consonant errors

1. If aspirated and unaspirated consonants can be interchanged then interchange it for the standardized spelling. For example,
    a. -b (ब) and -b^{h} (भ)
        i. sabi (सबी) -> sab^{h}i (सभी)
    b. -t (ट) and -th (ठ):
        i. meeta (मीटा) -> meet^{h}a (मीठा)
    c. -j (ज) and -jh (झ):
        i. janDaa (जंडा) -> j^{h}anDaa (झंडा)
    d. -g (ग) and -g^{h} (घ)
        i. gar (गर) -> g^{h}ar (घर)F
    e. -k (क) and -KH (ख)
        i. d^{h}oka (धोका) -> d^{h}ok^{h}aa (धोखा)
    f. -f (फ़) and -p^{h} (फ):
        i. fool (फ़ूल) -> p^{h}ool (फूल)
        ii. pheslaa (फैसला) -> feslaa (फ़ैसला)

2. If consonants are spoken in a way that is very close to the correct form (can be common or uncommon in spoken form) then change it to the correct form. For example,
    a. -s and -sh
        i. siksha(सिक्षा) -> shikshaa (शिक्षा)
    b. -d`h and - r`h
        i. pad`hnaa (पढना) -> par`hnaa(पढ़ना)
    c. -sh and -s :
        i. s^{h}os^{h}al (शोशल) -> sos^{h}al (सोशल)
        ii. saamil (सामिल) -> s^{h}aamil (शामिल)
        iii. chasmaa (चस्मा) -> chas^{h}maa (चश्मा)
    d. -d` and -r
        i. larka(लरका) -> lad`ka (लड़का)
        ii. b^{h}eed` (भीर ) -> b$^{h}$eer` (भीड़)
        iii. bar`tan (बड़तन) -> bartan (बर्तन)
        iv. sad`ak (सडक) -> sar`ak (सड़क)
    e. Interchange between -cch and -ksh
        i. ikshaa (इक्षा) -> icchaa (इच्छा)
        ii. cchetra (छेत्र) -> kshetra(क्षेत्र)

3. If the interchange of consonant clusters helps in achieving the standard spelling for a word (metathesis), such a transformation is allowed as long as it does not deviate the resultant word from the spoken-text. For example,
    a. Interchange between -tf and -ft
        i. luft (लुफ़्त) -> lutf (लुत्फ़)
        ii. Ustah (उस्ताह) -> Utsah (उत्साह)

4. If consonants are mispronounced because of its adjoining letter, then correct it (phonological assimilation). For example,
    a. Kuut.ta (कूत-ता) -> Kuud.ta (कूदता)
    b. kitap parh ke (किताप पढ़ के) -> kitab parh ke (किताब पढ़ के))
    c. kus sunna (कुस सुना) -> kuch sunna (कुछ सुना)
    d. chammas se (चम्मस से) -> chammach se (चम्मच से)

[10] https://ai4bharat.iitm.ac.in/indicvoices

Rules for correcting vowel errors

1. If vowel sounds are elongated or shortened or interchanged due to emphasis/modulation then change it to the standard form. For example,
    a. -a and -aa
        i. kaahaanii (काहानी) -> kahaanii (कहानी)
        ii. bahaar (बहार) -> baahar (बाहर)
    b. -i and -ii :
        i. miisaal (मीसाल) -> misaal (मिसाल)
    c. -u and -uu :
        i. uudhaar (ऊधार) -> udhaar (उधार)
        ii. ungli (उंगली) -> uungli (ऊँगली)
        iii. uncha(उंचा) -> uuncha (ऊँचा)
    d. -o and -u :
        i. Kyunki (क्यूँकि) -> kyonki (क्योंकि)
    e. -e and -a:
        i. bechen (बेचेन) -> bechain(बेचैन)
        ii. chain (चैन) -> chen (चेन) (गले की चेन)
    f. -a and -i :
        i. kataab (कताब) -> kitaab (किताब)
        ii. salaaii (सलाई)-> silaaii (सिलाई)

2. If interchanging monothongs (a, aa, e, i, ii, o, u, uu) with dipthongs (ai, au, etc) allows you to arrive at the standard spelling for the word, then please interchange. Note that diphthongs and triphthongs are combinations of two or three vowels respectively (aa + o = au). For example,
    a. hei(हे) -> hai (है) [e v/s ae]
    b. mei (में) -> mein (मैं) [e v/s ae]
    c. nao (नो) -> nau (नौ) [o v/s au]

3. If removal or addition of a vowel helps in getting to the standard spelling of a word, such a transformation is allowed. For example,
    a. removal of -a (using halant)
        i. darad (दरद) -> dard (दर्द)
        ii. mulak (मुलक) -> mulk (मुल्क)
        iii. khatam (खतम) -> khatm (खत्म)
    b. removal of -i (using halant) :
        i. Kirya (किरया) -> Kriya (क्रिया)
    c. addition of -i
        i. Briyani (ब्रियानी) -> biryani (बिर्यानी)

4. If vowels are replaced by consonants because of accent, then change it to the standard form. For example,
    a. -gu and -u :
        i. gunnees (गुन्नीस) -> unnees (उन्नीस)
    b. -a and -h:
        i. haspataal (हस्पताल) -> aspataal (अस्पताल)

Rules for dealing with errors at word beginnings

1. If word beginning requires insertion/deletion of a vowel or a schwa, then change it to the standard spelling. For example,
    a. Deletion of इ:
        i. Istree (इस्त्री) -> Stree (स्त्री)
        ii. Isthan (इस्थान) -> स्थान (Sthan)

Rules for dealing with errors at word endings

1. If word ending has an extra vowel then correct it to make it to standard form. For example,
    a. Himalayaa(हिमालया) -> Himaalya (हिमालय)
    b. kavii (कवी) -> kavi (कवि)
    c. matii (मती) -> mati (मति)

2. If additional schwa is added at the end of words ending in consonant sounds then remove it to bring it to standard form. For example,
    a. vageraa (वगैरा) -> vagerah (वगैरह)
    b. yogaa (योगा) –> yoga (योग)
    c. karmaa (कर्मा) -> karma (कर्म)
    d. raamaa(रामा) -> raam(राम)

3. If a consonant is missed or skipped at the end of a word then add it to bring it to the standard form. For example,
    a. missed -ta
        i. takh (तख) -> taKHt (तख्त)
    b. missed -h:
        i. Guna (गुना) -> gunah (गुनाह)
        ii. fate (फते) -> fatah (फतह)

Rules for dealing with shortened syllables

1. If splitting of a shortened syllable followed by character additions standardizes the spellings of the words, then such a transformation should be allowed, while ensuring the sound of the resultant words do not deviate from the spoken text
    a. Bolraa (बोलरा) -> Bol rahaa (बोल रहा)
    b. Jaraa (जारा) -> Ja rahaa (जा रहा)
    c. Manraa (मानरा) -> Maan rahaa (मान रहा)
    d. Mujhebi (मुझेबी) -> Mujhe bhi (मुझे भी)
    e. Muje-i (मुजेई) -> Mujhe hi (मुझे ही)
    f. Teraa-i (तेराई) -> teraa hi (तेरा ही)

2. If present tense markers have been shortened or eliminated in the phrases, then change them to arrive at the standard form.
    a. shortening of hai (है), hun (हूँ) and hain(हैं) :
        i. Kartau (करताऊँ) -> Karta hun (करता हूँ)
        ii. Kartaye (करताए) -> Karta hai (करता है)
        iii. Aatain (आतेए) -> Aate hain (आते हैं)

Rules for dealing with nasalised sounds

1. If addition or deletion of a nasalised sound can bring a word into its standard form, then such a transformation is allowed if it does not impact the context in which the word was uttered. e.g.:
    a. addition of diacritic (बिंदु - ं and चंद्रबिंदु ँ)
        i. mujhe kae jana tha (मुझे कही जाना है) -> mujhe kahin jaana tha (मुझे कहीं जाना है)
        ii. ye saare mere ghanisht mitr hai (ये सारे मेरे घनिष्ठ मित्र है) -> ye saare mere ghanisht mitr hain (ये सारे मेरे घनिष्ठ मित्र हैं)
        iii. kahaaniyaa (कहानिया) -> kahaaniyaan (कहानियाँ)
        iv. yaha (यहा) -> yahaan (यहाँ)
    b. deletion of (बिंदु) :
        i. mere kaen dost hain (मेरे कईं दोस्त हैं) -> mere kayi dost hain (मेरे कई दोस्त हैं)
    c. deletion of unnecessary '-n' sound:
        i. Karenga (करेंगा) -> Karega (करेगा)
        ii. Bolenga (बोलेंगा) -> Bolega (बोलेगा)
        iii. Jaayenga (जाएँगा) -> Jaayega (जाएगा)

Rules for dealing with loan words

1. If common words borrowed from english or english named entities such as proper nouns, places, brands, food, month/day names, etc are present in isolation without any inflection (-ein, -ne, -ka, etc), then add the english spelling in brackets, while leaving the original transcript as verbatim. For example,
    a. Proper Nouns
        i. amjon (ऐमजॉन) -> (ऐमजॉन) [amazon] (note that the (हिंदी) spelling has **not** been corrected from "amjon " to "amazon" but the correct english spelling has been added in brackets []
        ii. aadidas (आदिदास) -> आदिदास [adidas]
        iii. aktubar(अक्टूबर) -> अक्टूबर [october]
        iv. wednasde (वेन्सडे) -> वेन्सडे [Wednesday]
        v. disember(दिसंबर) -> दिसंबर [December]
    b. Common words
        i. on (ओन) -> ओन [on]
        ii. ooder (ओडर) -> ओडर [order]
        iii. order(ओर्डर) -> ओर्डर [order]
        iv. fusT(फस्ट) -> फस्ट [First]
        v. baseball(बेसबॉल) -> बेसबॉल [baseball]
        vi. Cupcake(कपकेक) -> कपकेक [Cupcake]
        vii. Playground(प्लेग्राउंड) -> प्लेग्राउंड [playground]
        viii. kilogram(किलोग्राम) -> किलोग्राम [kilogram]

2. If certain sounds in borrowed words are not present in the language's orthography (names/places from different cultures, states) then come up with the closest way as mentioned in the conventions of your language and use that consistently across all instances of the same word. The same holds true for words which can be written in multiple ways. The idea is to ensure consistency across all utterances of a spoken word.
    a. tamiḷ (तमिळ) -> tamil (तमिल)

3. If acronyms are spoken word by word (in English) then always write it with a space. E.g.:
    a. NRI -> N R I (एन [N] आर [R] आई [I])
    b. FBI -> F B I (ऐफ़ [F] बी [B] आई [I])
    c. MPin -> M Pin (एम [M] पिन [Pin)
    d. SBI -> S B I (एस [S] बी[B] आई [I])

4. If acronyms are spoken as a single word then write it as it is because spacing it would deviate it from the spoken-text. E.g:
    a. ASAP (एसेप) and not A S A P (ए [A] एस [S] ए [A] पी [P])
    b. BhaJPa (भाजपा) and not B J P (बी जे पी) or Bha J Pa (भा जे पा)

Rules for dealing with numbers, units of time, measurement, etc

1. If numbers (ordinals and cardinals), time, and other units of measurement (kg, m, cm, ft, etc) are uttered in English then represent them exactly as it is in the target transcription language. E.g.:
    a. kilogram (किलोग्राम [kilogram])
    b. kg (कि.[k] ग्रा.[g])
    c. first (फर्स्ट[first])

2. If numbers (ordinals and cardinals), time, and other units of measurement (kg, m, cm, ft, etc) are uttered in the native language, then always use the standardized native language norms and spellings. e.g.:
    a. gunniis (गुन्निस) -> unniis (उन्नीस)
    b. paach (पाच) -> paanch (पाँच)
    c. miTer (मिटर) -> miiTer (मीटर)
    d. pelii (पेली) -> pahli (पहली)

Rules for compound/conjunction/combined words/Reduplication

1. If compound words are used in non-standard form, then conjunction can be added to change it to the standard form.
    a. Inapplicable for Hindi

2. In case of reduplicated words, when the next word is not in the dictionary, please add a hyphen(-) in between and write it as it is spoken.e.g:
    a. chai vai (चाय-वाय)
    b. Roti voti (रोटी-वोटी)
    c. daal vaal (दाल-वाल)

3. If insertion or removal of a space is required to standardize the spelling of a compound word, such a transformation is allowed.
    a. Inapplicable for Hindi

Rules for word contractions:

1. Inapplicable for Hindi

Rules for colloquial tardiness:

1. If addition of a genitive marker results in standard form, then such a transformation is allowed.
    a. addition of -ri (री), -ra (रा), -re (रे)
        i. humaai kitab (हमाई किताब) -> humari kitaab (हमारी किताब)
        ii. humaye gaaon me (हमाए गाँव में) -> humare gaaon me (हमारे गाँव में)

    Note that here both “humaai”(हमाई) and “humaye” (हमाए) are not being changed to “hum” (हम) because it differs significantly from the spoken text.

2. If changes in dative (direction) forms help in arriving at a standard spelling, then such a transformation is allowed to a level where it does not deviate much from the spoken text:
    a. tumar ko (तुमार को) -> tumhare ko (तुम्हारे को)
    b. humar ko (हमार को) -> humare ko (हमारे को)
    c. mere ko (मेरे को) -> mere ko (मेरे को) [no change]
    d. mer ko (मेर को) -> mere ko (मेरे को)
    e. meko (मैंको) -> main ko(मैं को) [split to meaningful units]
    f. tumaare ko (तुमारे को) -> tumhare ko (तुम्हारे को)

    Note: “tumar” (तुमार) is not being changed to “tum” (तुम), “mere” (मेरे) is not being changed to “mujhe” (मुझे) because they differ significantly from the spoken text.

3. If a small change in form of a word helps in arriving at a correct spelling, then such a change is allowed.
    a. Bol riya (बोल रिया) -> bol rahaa (बोल रहा)
    b. Bol riye (बोल रिए) -> bol rahe (बोल रहे)

Rules for gemination:

1. The gemination (doubling consonant) is represented in orthography with diacritics, matras or by doubling consonants. Please follow the conventional way of writing geminated words in your language and maintain the consistency across all spellings of an utterance. e.g.
    a. Addition of halant (`)
        i. patta (पत.ता) --> patt+ta (पत्+ता -> पत्ता)
        ii. kaccha (कच.चा) -> kach+cha (कच्+चा -> कच्चा)

2. If double emphasis in words is not very obvious but the speaker means it, then change it to the double emphasis form. This could happen in instances of proper nouns with no standardized spellings, brand names, lesser known food items, brands, etc. e.g.:
    a. apam (अपम) -> appam (अप्पम)
    b. Aka (अका)-> akka (अक्का)
    c. Ana (अना) -> Anna (अन्ना)
    d. bhayaa (भया) -> bhayyaa (भय्या)

Don’t change during Level-2:

1. Borrowed words with native suffixes (mostly plural) should be written as it is, without changing or translating the suffixes. e.g.:
    a. calendar-on (क्लैंडरों) should not be written as calendar (कैलेंडर)
    b. side-ein (साइडें) should not be written as side (साइड)
    c. Kantriyoun (कंट्रीयों) should not be written as countries (कन्ट्रीज)
    d. Friendo (फ्रेण्डों) should not be written as friends (फ्रेंड्ज़)

2. If splitting of a shortened syllable followed by character additions standardises the spellings but makes it deviate significantly from the waveform (spoken text), then we should not be allowing such transformation:
    a. unne (उनने) <-> unho ne (उन्होंने) [not-allowed]
    b. inne (इनने) <-> inho ne (इन्होंने) [not-allowed]

3. Spoken/dialectic words should not be replaced by the standard versions, as they deviate a lot from the wave form.
    a. aho (आहो) <-> haan (हाँ) [not-allowed]
    b. Hav (हव) <-> haan (हाँ) [not-allowed]
    c. abe oye (अबे ओय) <-> arey o (अरे ओ) [not-allowed]

4. If new affix patterns, echo words or reduplication patterns are innovated by the speaker, or slangs used only in spoken languages are found, then retain them as it is (with space) as
   corrections will deviate too much from the spoken form. E.g:
   a. Loot laat (लूट लाट) [shall be retained as it is]
   b. Khaali peeli (खाली पीली) [shall be retained as it is]
   c. Paani puuni (पानी पूनि) [shall be retained as it is]

5. If grammatical corrections make the resultant text deviate much from the spoken-text, then such changes should not be made and original words should be retained. e.g.:
   a. Corrections disallowed in dative forms
      i. mereko ghar jana hai (मेरे को घर जाना है) <-> mujhe ghar jana hai (मुझे घर जाना है) [not-allowed]

6. The conjunction should be infixed, or removed only where it's absolutely necessary as per standardized spelling rules. If the 2 words can stand alone and have a spelling then, do not add meaningful affixes as per the rule.
   a. Inapplicable for Hindi

7. Severe colloquial contractions in Tamil and other languages, such as "vandethanalla" should not be expanded to their textbook forms such as"vande vithana allava"
   a. Inapplicable for Hindi

[End]