Appendix: INDICVOICES-R: Unlocking a Massive Multilingual Multi-speaker Speech Corpus for Scaling Indian TTS

A Datasheets for Datasets

The following section is answers to questions listed in datasheets for datasets

A.1 Motivation

- **For what purpose was the dataset created?**
  INDICVOICES-R is created to scale-up Indian TTS to all 22 languages, covering a large number of speakers from various demographics. We also release INDICVOICES-R-Benchmark to evaluate the cross-speaker generalization of TTS systems for Indian voices spanning across age-groups and real life scenarios.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  INDICVOICES-R is presented by AI4Bharat, a research lab at the Indian Institute of Technology-Madras, whose mission is to empower and elevate the potential of Indian languages in AI technologies by fostering open-source collaboration in datasets, models, and applications.

- **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.**
  The dataset creation was funded by Digital India Bhashini, the Ministry of Electronics and Information Technology of the Government of India, EkStep Foundation and Nilekani Philanthropies.

A.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries**
  INDICVOICES-R contains speech-text pairs with additional metadata such as SNR, C50, speaking rate, gender, age-group, etc.

- **How many instances are there in total (of each type, if appropriate)?**
  INDICVOICES-R comprises a total a 1704.34 hours of speech-text pairs from approximately 690K utterances, containing 10,496 speakers across the 22 officially recognized languages of India. Please refer Table 3 for more details.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
  Yes, this dataset is a complete set. Although the dataset is derived from IndicVoices, it is complete in itself and is intended to be used for TTS (Text-to-Speech) research and applications.

- **What data does each instance consist of?**
  The dataset contains speech-text pairs i.e., the filepath pointing to each audio file and the corresponding normalized text. Each instance also includes metadata on the speech from metrics like SNR, C50, pitch mean and speaking rate to metadata on the speaker like age-group, gender, etc.

- **Is there a label or target associated with each instance?**
  Yes, there are text, gender, speaker and language labels associated with each instance of the data.

- **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.**
  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
  Yes, instances containing audio recordings from the same speaker are identified with the anonymized speaker ids, which is captured in the meta-data.

- **Are there recommended data splits (e.g., training, development/validation, testing)?**
  Yes, we provide the training and validation splits to ensure that there's no test-set leakage from the benchmark. We release this on our dataset webpage `https://github.com/AI4Bharat/IndicVoices-R`.

- **Are there any errors, sources of noise, or redundancies in the dataset?**
  No, the data does not have any errors, sources of noise or redundancies.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
  The dataset is self-contained and complete in itself.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?**
  No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
  No.

- **Does the dataset relate to people**
  Yes, this dataset consists of audio recordings from multiple people but is derived from a source which has already ensured anonymization.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**
  Yes, the dataset provides desensitised metadata on age-groups, gender and more.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
  No, the data has been thoroughly anonymized, ensuring that identification of individuals is not possible with the provided metadata.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
  The data does not contain any sensitive metadata as such, but includes metadata from its parent source such as age-group, gender, occupation, and location, which have all been thoroughly desensitised.

A.3 Collection Process

- **How was the data associated with each instance acquired?**
  We derive our dataset by enhancing speech-text pairs from the existing IndicVoices dataset, followed by careful filtering. The entire process has been discussed in Section 3 of the paper.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**
  Our work transforms an existing dataset using several algorithms that have been discussed in Section 3.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  N/A.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
  This dataset was created by students and project staff who were paid standard industry rates.

- **Over what timeframe was the data collected?**
  N/A.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
  An ethical review was conducted for the parent dataset which deemed the dataset to be fit for use for any downstream applications.

- **Does the dataset relate to people?**
  Yes, this dataset consists of audio recordings from multiple people but is derived from a source which has already ensured anonymization.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  The parent data was acquired from IndicVoices (`https://ai4bharat.iitm.ac.in/indicvoices/`) which is a publicly released ASR dataset spanning across 22 languages.

- **Were the individuals in question notified about the data collection?**
  N/A.

- **Did the individuals in question consent to the collection and use of their data?**
  N/A. We re-purpose an existing dataset released under the CC-BY-4.0 license comprising of speech-text pairs for which explicit consent has already been collected.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
  N/A.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
  N/A.

A.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  Our dataset is derived from an existing ASR dataset, and we thoroughly describe the steps of data preparation, filtering, and post-processing in Section 3.2.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

N/A. The source of our parent dataset is publicly available.

- **Is the software that was used to preprocess/clean/label the data available?**

  1. **Preprocessing:** The data was made fit to be use for TTS systems by re-purposing the ASR data, IndicVoices through a cascade of systems:
     - HTDemucs (https://github.com/ZFTurbo/Music-Source-Separation-Training)
     - VoiceFixer (https://github.com/haoheliu/voicefixer)
     - DeepFilternet-3 (https://github.com/Rikorose/DeepFilterNet/)
  2. **Cleaning and Filtering:** Data stats such as C50, SNR, speaking rate, pitch stats, were computed using DataSpeech (`https://github.com/huggingface/dataspeech`) and NORESQA was calculated using the official implementation (`https://github.com/facebookresearch/Noresqa`).

A.5 Uses

- **Has the dataset been used for any tasks already?**
  Yes, in this work we use INDICVOICES-R to train the first ever TTS model covering all 22 Indian languages and we report its performance scores on our proposed benchmark in Table 5.

- **Is there a repository that links to any or all papers or systems that use the dataset?**
  N/A.

- **What (other) tasks could the dataset be used for?**
  The dataset could be augmented and used for the several purposes its parent dataset boasts of including speaker diarization, speaker identification, speaker verification, language identification, intent detection, entity extraction, query by example, audio denoising, speaker re-construction, and speech separation.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  No.

- **Are there tasks for which the dataset should not be used?**
  No, but the authors call for responsible usage of the dataset.

A.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  The dataset is publicly available for everyone to use under the said license.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
  The dataset is released at `https://github.com/AI4Bharat/IndicVoices-R` and comprises of tarballs split across languages which can be downloaded.

- **When will the dataset be distributed?**
  13/06/2024 and onwards.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
  The dataset is released under CC-BY-4.0 License.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
  No.

A.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  AI4Bharat, our research lab at IIT Madras, will support, host, and maintain this dataset.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  Please contact the authors or the provider, AI4Bharat, through Github issues at `https://github.com/AI4Bharat/IndicVoices-R`.
- **Is there an erratum**
  No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
  Any updates or corrections if required will be reflected on the official GitHub page.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
  N/A.

- **Will older versions of the dataset continue to be supported/hosted/maintained?**
  We are currently at the first version of our dataset and intend to support all versions even during future releases.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
  We are open to discussing such requests and would kindly ask users to contact the authors for the same.

B Dataset Nutrition Label

Variables	
text	Level 2 transcript. The text is normalized for abbreviations, numbers and standardized spellings. Please refer to IndicVoices for details on the normalization rules.
lang	Language of the utterance, listed in the ISO-639-1 format.
samples	Number of samples = Duration of audio $\times$ *sampling_rate*.
verbatim	Level 1 transcript. The text captures the spoken language as it is, without any standardization of spellings.
normalized	Level 2 transcript. The same as the field *text*.
speaker_id	The ID of the speaker in the utterance.
scenario	The speech may be Read-Speech or Extempore. In case of Extempore, the speakers were not provided with a script, instead they had to answer according to a prompt or question.
task_name	Represents the domain to which the utterance belongs. The dataset covers more than 70 domains, ranging from styles like Alexa commands to topics like Games, Tourism and Technology.
gender	The participants were asked to indicate their gender, with the options 'Male', 'Female' and 'Others'.
age_group	The data has speakers from four age groups, 18-30 years, 30-45 years, 45-60 years and 60+ years.
job_type	Participants are classified into four job categories - 'Student', 'Unemployed', 'Blue Collar', 'White Collar'.
qualification	Participants could choose from four qualification types - 'No Schooling', 'Post-Grad + PhD', 'Upto 12th', 'Undergrad and Grad'.
area	The place of recording - Rural or Urban.
district	The district within the state in which the utterance was recorded.
state	The state of India in which the utterance was recorded.
occupation	Participants were requested to declare their occupation. Our dataset covers utterances from all walks of life.
filename	Name of the audio file corresponding to the utterance.
duration	Duration of the utterance, reported in seconds.
cer	This is the Character Error Rate between Level 1 and Level 2 transcripts.
snr	The Signal-to-Noise ratio calculated using Brouhaha.
C50	The reverberation score, C50, calculated using Brouhaha.
utterance_pitch_mean	The mean of the pitch within the utterance computed using PENN.
utterance_pitch_std	The standard deviation of the pitch within the utterance computed using PENN.

C Qualitative Examples from the INDICVOICES-R Benchmark

Language Code	Scenario	Sentences
asm	Alexa Commands	**Text:** এটা ষ্টটোছ দয়া. এইটো কৈ মই প্ৰমোচন পাইছোঁ **Transliteration:** atta status diya eitu koi moi promuson paisu **Translation:** Put a status saying 'I got a promotion'.
ben	Domain of Interest - Agriculture	**Text:** তাতে একটু সময় বাঁচলো মানুষের আর অল্প সময়ের মধ্যে অনেক জমি জায়গা চাষবাস করা হতো **Transliteration:** tate ektu somomo banchlo manusher aar alpo somepor modhye onek jommy jagiba chashbas koraa hoto **Translation:** This saved time for the people and in a short time, a lot of land was cultivated.
brx	Umang Commands	**Text:** केराला रायजोआव स्पुटनिक गोनां गासैबो सेन्टारफोरखौ आंनो दिन्थि **Transliteration:** kerala raijwao spootnik gwnang gaswibw cen-trepwrkwo angnw dinti **Translation:** Show me all centres in state Kerala having Sputnik
doi	Domain of Interest - Animal Husbandry	**Text:** डेरी आह्ले जेह्ड़े फार्मर होंदे दुद्ध शुद्ध आह्ले ते ओह् ते गमां ते मंजां पालदे **Transliteration:** dairy ahle jehde former honde duddh shuddh ahle the oh the gamaan the manjaan paalade **Translation:** Farmers like the dairy farmers would raise milk and they would raise cows and cattle
guj	Bigbasket Commands	**Text:** કેસરી સેફ્રોનને પરત કરવાની શું પ્રક પ્રક્રિયા છે બરાબર **Transliteration:** kesari sephrone parat karvaanee shun prak prakriyaa chhe barabar **Translation:** What is the procedure to return Kesari Saffron?
hin	Domain of Interest - Business	**Text:** बहुत सारे ऐसे बिजनेस हैं जैसे कि कपड़ों का बिजनेस है **Transliteration:** bahut saare aise business hain jaise kii cups kaa business hai **Translation:** There are many such businesses like clothing business.
kan	Know your participant - Traveling	**Text:** ಎಲ್ಲರು ಮಾಸ್ಕ್ಗಳನ್ನು ಕಡ್ಡಾಯ ಹಾಕಬೇಕಂತಿತ್ತು ಮತ್ತೆ **Transliteration:** ellaru maskgalannu kaddaya haakabekantittu matte **Translation:** It was mandatory for everyone to wear masks again
kas	Daily Life	**Text:** ھُچ یرأخبُ ھُچ رٹیٖہ ھُچ ہِی زٔنٛم سنَچِک ھُچ سنَیٚکنٟو **Transliteration:** venkenas chu kichnas menz yeh chu heater chu bukheyr chu **Translation:** Currently in the kitchen, it is hotter than usual.
kok	Coordinator Section	**Text:** तुमच्या मते नितळसाणीचें कितें म्हत्व आसा **Transliteration:** tumchya matte nitalsaanichen kitein mhatva aasaa **Translation:** What do you think is the importance of cleanli-ness?
mai	Keywords Spotting	**Text:** असहमति **Transliteration:** asahamati **Translation:** Disagreements
mal	Domain of Interest - Health	**Text:** ഞങ്ങളുടെ ആരോഗ്യകേന്ദ്രമെന്ന് പറയുമ്പോൾ ഇവിടെ ഇവിടെ ഹോസ്പിറ്റലുണ്ട് ജില്ലാ ആശുപത്രി **Transliteration:** njangalude aarogyakendramennu parayumbol ivide ivide hospitalundu jilla ashupatri **Translation:** When we say our health center, there is a hospital here, the district hospital

mni	Domain of Interest - Culture	**Text:** ꯄꯤꯗꯨꯅ ꯂꯥꯛꯂꯤ **Transliteration:** piduna laakli **Translation:** It is being supplied.
mar	Domain of Interest - Entertainment	**Text:** राजकारणामध्ये सुद्धा ते भाग घेतात समाजसेवा करतात असा असा वेळ तो आपला फावल्या वेळामधे आपला करतात आणि ये **Transliteration:** rajkaranamadhye suddha they bhag ghetat samajseva karatat assa assa vela too aapala favalya velamadhe aapala karatat aanee yeey **Translation:** He also participates in politics and does social service in his spare time.
nep	Domain of Interest - Animal Husbandry	**Text:** के रे दुधहरू दुहुनु जाँदाखेरि कति गाईले त लात हान्ने गर्छ कति गाईले **Transliteration:** key ray dudhaharu duhunu jandakheri kathi gaile tata lat hanne garchha kathi gaile **Translation:** How many cows kick while going to milk milk?
ori	Domain of Interest - Legal	**Text:** ତେଣୁ ଯଦି ଶିଶୁମାନେ ଶ୍ରମିକ ହୋଇଯିବେ ତାହେଲେ ଦେଶ ବୁଡ଼ିଯିବ **Transliteration:** tenu jadi shishumane shramika hoijibe taahele desh bujiba **Translation:** So if children become labourers, the country will sink.
pan	Domain of Interest - Health	**Text:** ਪੰਜਾਬ ਦੇ ਪ੍ਰਸ਼ਾਸਨ ਵੱਲੋਂ ਸਹਿਤ ਸੰਭਾਲ ਉੱਤੇ ਬਹੁਤ ਖਾਸ **Transliteration:** punjab dey parshasana vallon sihat sambhal utte bahut khaas **Translation:** Punjab's administration is very particular on healthcare
san	Domain of Interest - Religion	**Text:** विघ्ननाशाय गणहोमं वा ऐश्वर्य ऐश्वर्यादिप्राप्त्यर्थं लक्ष्मी उपासना वा **Transliteration:** vighnanaashaaya ganahomam wa aishvarya aishwaryaadipraaptyartham lakshmi upasana wa **Translation:** Vighnashay Gana Homam or Lakshmi Puja for wealth and prosperity.
sat	Domain of Interest - Geography	**Text:** ᱠᱚᱨᱟ. ᱤᱮ ᱵᱟᱤᱟᱝ ᱯᱮ ᱫᱚᱨ-ᱟ ᱟᱤ ᱪᱩᱱᱟ. ᱞᱟᱹ ᱳᱮ ᱵᱟᱤᱟᱝ ᱫᱚᱮᱳᱚ ᱦᱩᱭᱩᱜ-ᱟ **Translation:** Karam is performed during the night and Karam is performed along with the puja.
snd	Digital Payment Commands	**Text:** छा तव्हां जिओ पे तां मुंहिंजो नैनीताल बैंक मास्टरकार्ड हटाए सघंदा आहियो **Transliteration:** chhaa tavhaan geo pay taan munhinjo nainital bank mastercard hataaye saghanda aahiyo **Translation:** Can you remove my Nainital Bank Mastercard from Jio Payments Bank?
tam	District Specific	**Text:** இந்த மதிரிஅழகர் திலகர் திடல் **Transliteration:** inththa madhiri azhagar thilagar thidal **Translation:** This is the type of beautiful tilak.
tel	Know your participant - Basic	**Text:** సొంతూరు అంటే చాలా ఇష్టం అన్నమాట **Transliteration:** sonturu ante chaalaa ishtam annamata **Translation:** I love being at home
urd	District Specific	**Text:** چھوٹے بچوں کا **Transliteration:** chhote bachon kaa **Translation:** the little children's

D Authors' Statement, Ethics, and Privacy

In this work, we introduce a transformed dataset derived from the previously publicly released IndicVoices dataset. We re-release our transformed dataset under the same CC-BY-4.0 license to ensure its continued accessibility and utility for the research community. The parent dataset's data collection process underwent a rigorous ethical review and approval by the Institute Ethics Committee, ensuring comprehensive ethical standards were upheld. This included providing all instructions in participants' native languages, fully informing participants about the purpose of data collection, obtaining explicit consent, and offering appropriate compensation. We have adhered to the same ethical standards in our work, safeguarding the privacy and confidentiality of the participant's personal identifiable information (PII), and ensuring the data remains anonymized and protects sensitive information. All personnel involved in the transformation and handling of the dataset were appropriately compensated, maintaining ethical standards in labor practices. By releasing this transformed dataset under the CC-BY-4.0 license, we support its free and open use, including for commercial purposes, in line with the original dataset's licensing terms. We commit to updating our dataset in accordance with any changes to the license terms of the original dataset, ensuring continued compliance and ethical integrity.

E DOI

The Digital Object Identifier for the IndicVoices-R is 10.5281/zenodo.11636051.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Section 4, 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Section 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see Section 5
   (b) Did you mention the license of the assets? [Yes] see Section 6
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] see Section 1
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] see Section 6
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] see Section 6

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]