german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

BIZAUGMENTOR

FAIR FORWARD
Artificial Intelligence for all.

A Study on
Open Voice Data in Indian Languages

Published by:
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn

A Study on Open Voice Data in Indian Languages
A2/18, Safdarjung Enclave
T: +91 11 4949 5353
F: +91 4949 539
E: fairforward@giz.de
I: www.giz.de/india

Person responsible
Mr. Gaurav Sharma
E: gaurav.sharma1@giz.de
FAIR Forward – Artificial Intelligence for All

Author:
BizAugmentor Global Services Pvt Ltd
www.bizaugmentor.com

Design and Layout
Caps & Shells Creatives Pvt Ltd

Photo credits
All the images that have been used in the report are openly accessible and made to use from 'Google' with permission for reusability and modification.

On behalf of the
German Federal Ministry for Economic Cooperation and Development (BMZ)

GIZ is responsible for the content of this publication
New Delhi, December 2020

Preface

The language diversity and lack of technological support for spoken languages in India make universal access to information and services an ongoing challenge. AI-based voice recognition has great potential. It makes technology more inclusive and enables millions of people to access services they cannot use yet – be it in agriculture, education, health, or others. For this vision to be realised, a central obstacle needs to be addressed: The lack of free and open voice data in India to develop and train natural language processing models.

The present research work is conducted as part of the global project 'FAIR Forward – Artificial Intelligence for All', initiative of the German Federal Ministry for Economic Cooperation and Development (BMZ), implemented in India by Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Gmbh. This initiative is also active in partner countries: Ghana, Rwanda, South Africa and Uganda.

This research aims at finding the current state of (open) voice datasets in Indian languages, including information about their volume, quality, mode of collection, and availability. The present research also explores the challenges for the creation and maintenance of open voice datasets in India. The report makes practice-oriented recommendations for future sustainable voice data collection based on both extensive desk research and expert interviews.

Contents

LIST OF TABLES

LIST OF FIGURES

Executive Summary

The size of the global smart speaker market was valued at $4.3 billion in 2017, which increased to $8.4 billion in 2019[23]. This market is expected to grow further at a CAGR of 17.1 percent until 2025. The majority of this growth has been contributed by North America, Europe, and other countries where English is the most common communication language.

Artificial intelligence (AI) and speech systems have amplified human effectiveness by augmenting their capability to access information and to interact with machines. These systems have increased human productivity in day-to-day life and automated numerous manual tasks in industrial scenarios. Home assistance devices like Amazon Echo or Google Nest help people manage their homes more effectively.

The size of the global smart speaker market was valued at $4.3 billion in 2017, which increased to $8.4 billion in 2019[23]. This market is expected to grow further at a CAGR of 17.1 percent until 2025. The majority of this growth has been contributed by North America, Europe, and other countries where English is the most common communication language. India being the second

largest English-speaking nation in the world, has also contributed significantly to this growth. However, merely 10-12 percent of Indian population speaks English.

The potential and opportunity that AI offers to the Indian society and economy have not been explored in-depth. AI products and services have not been able to reach significant sections of Indian society since India is home to thousands of languages (mother tongues), and people in every state (or regions within a state) communicate in their native language (or dialect). If AI speech systems are to reach every section of Indian society, they must incorporate the variations and dialects of regional Indian languages.

There are large amounts of open voice datasets available for the English language, which are enough to develop good quality AI speech systems; however, similar assurance cannot be given for speech datasets in Indian languages. If enough volume of good quality (open) voice datasets could be made available for Indian languages, it can help create products and services that could be useful for people of different sections of the society. And can further help address many social issues related to education, healthcare, and many more. Therefore, it is imperative to explore and understand the availability of such resources for Indian languages and assess if they are suitable for developing AI-based voice recognition or speech synthesis systems.

This research focuses on exploring the availability of (open) speech datasets for Indian languages and listing their corresponding characteristics such as quantity, quality, etc. The report also elaborates on the properties related to the speech datasets and how they matter when collecting and applying such datasets for real-life applications. The report provides an overall picture of Indian language voice datasets' availability and also outlines the challenges and recommendations for scaling it up.

It is difficult to find a good quality general-purpose speech dataset for Indian languages that captures different dialects/accents associated with each language.

The datasets available with various institutes, researchers and organizations are usually domain-specific and not suitable for general-purpose speech systems. Moreover, the current efforts of data collection have been limited to covering the scheduled languages.

The data collected must follow a standard guideline that makes them suitable for reuse. A common platform needs to be built where the voice datasets from various sources can be collected and maintained in an organised manner. A standard agreement with a clear usage policy for data sharing also needs to be formulated.

The technology developed over voice datasets must be localised. Rather than taking people to a system, the system must be taken to people to have a social impact in India.

1. Languages in India

Language is a significant feature of any society. In India, language changes its features every few hundred miles. India speaks in a lot more languages than the number of languages for which there is available literature.

Occupying 2.4 percent of the world's land area and housing more than 18 percent of the world's population, India has 1369 classified mother tongues. Out of these classified languages (or mother tongues), 121 are spoken by 10,000 or more speakers. Twenty-two out of which are spoken by nearly 97 percent of the Indian population and are termed as scheduled languages. The remaining 99 non-scheduled languages account for the remaining 3 percent of the people[1].

The 22 scheduled languages put together, spoken by more than 1 billion Indians are listed below:

1. Assamese
2. Bengali
3. Bodo
4. Dogri
5. Gujarati
6. Hindi
7. Kannada
8. Kashmiri
9. Konkani
10. Malayalam
11. Maithili
12. Manipuri
13. Marathi
14. Nepali
15. Oriya
16. Punjabi
17. Sanskrit
18. Santhali
19. Sindhi
20. Tamil
21. Telugu
22. Urdu

In India, languages are broadly classified under five language families. Table 1 depicts the characteristics of each in terms of the population that recognises them as their mother tongue.

Table 1: Families of Indian languages along with number of speakers (Census of India, 2011)

Language families	Number of languages	People with language as their mother tongue	Approx. percentage of total population
Indo-European			
Indo-Aryan	21	945,052,555	78 %
Iranian	1	21,677	< 0.5 %
Germanic	1	259,678	< 0.5 %
Dravidian	17	237,840,116	20 %
Austro-Asiatic	14	13,493,080	1 %
Tibeto-Burmese	66	12,257,382	1 %
Semito-Hamitic	1	54,947	< 0.5 %
	121	**1,208,979,435**	

Out of these classified languages (or mother tongues), 121 are spoken by 10,000 or more speakers. Twenty-two out of which are spoken by nearly 97 percent of the Indian population and are termed as scheduled languages.

Most Indian people's mother tongue falls under the Indo-Aryan sub-family, with more than 945 million speakers.

numerous languages, each language has its own set of dialects. Each dialect is native to a region of the country (or state).

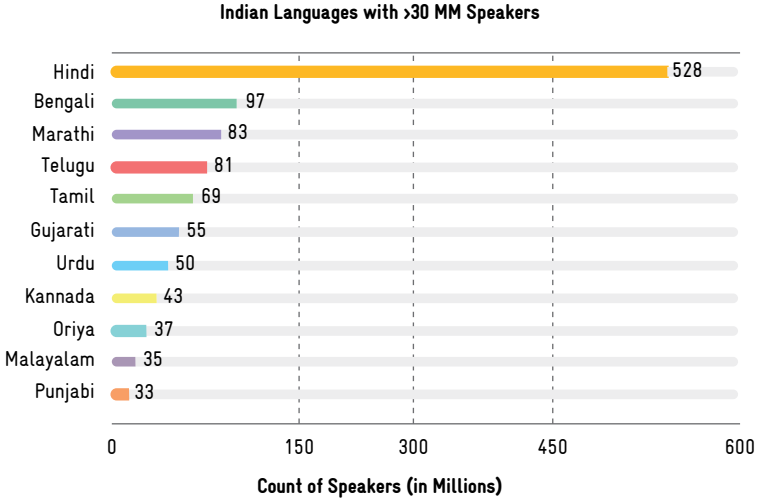


Figure 1: Top 11 Indian languages based on number of speakers (Census of India, 2011)

Among the Indo-Aryan family, Hindi and Bengali are the most widely spoken languages, and they are also among the top 10 most spoken languages of the world.

Based on the Indian Census of 2011, one can verify that Hindi is the most prominent language concerning the number of speakers, followed by Bengali, Marathi, Telugu, and Tamil.

Region-wise these languages broadly cover the country's landscape as depicted in figure 2, where they are spoken in large numbers. Apart from being a country with

The Hindi language has 48 officially recognised dialects[5]. For instance, in the eastern parts, Hindi has dialects such as Chhattisgarhi, Awadhi, Bagheli, and a few more, whereas the western regions have dialects such as Braj Bhasha, Hindustani, Bundeli, etc.

The variety of dialects can be found in almost all the major (scheduled) languages of the country. This variation makes the classification of languages and speakers based on states or regions even more challenging.

Figure 2: States and union territories of India by the most commonly spoken first language (50th Report of The Commissioner for Linguistics Minorities in India).

AI Industry in India: Focal Points

2.

In India, the number of internet users has increased by about 23 percent from January 2019 to January 2020 compared to the global growth rate of 7 percent. The majority of these users access the internet on their mobile phones. It has been found that around 90 percent of the Indian population in the 16 to 64 years age group owns a smartphone. As of January 2020, about 50 percent of the Indian people are mobile internet users.[7]

India is on the path to becoming a digital powerhouse, given the increased mobile phone usage and lower internet costs. As a result, the massive volume of data generated daily can potentially be leveraged by the AI community to build more accurate, real-world, and socially significant AI systems.

Over the last few years, AI has emerged as a powerful technology because of the availability of advanced algorithms and high computing power. Present-day AI algorithms retrieve information from various sources, analyze the data, and act according to insights derived from data.

The role of big tech companies like Google and Amazon in AI has mostly been in creating applications and platforms targeted towards India's middle and upper-middle-class population. This can be concluded from the fact that the language used in their products is mostly English (Hindi in some cases), which is widely spoken and read by middle and upper-middle-class people in India. Including other sections of society would also help strengthen these organizations' bottom line in the longer run.

Startups, on the other hand, often pursue a different business model. They strive for market presence to demonstrate exponential growth potential. They are known to value market reach more than profits. As the major proportion of India's population belongs to the country's rural and semi-urban regions, targeting them would become a promising strategy for such organizations to achieve significant growth.

The majority of startups in AI are into end-to-end analytics products and services (figure 3). They are primarily involved in providing solutions in terms of insights and future strategies for their clients (larger companies), based on historical performance data. Very few of them deal in NLP (Natural Language Processing) and speech related research and solutions. This limited focus on speech and NLP applications is mostly due to the dearth of good quality datasets in Indian regional languages that could help reach a significant section of the Indian population.

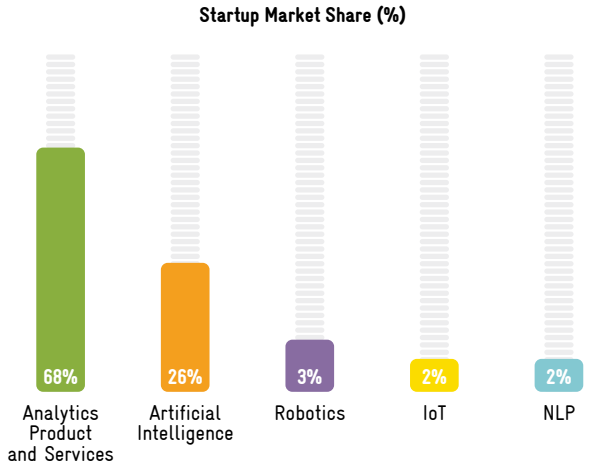


Figure 3: Startup market share (State of Artificial Intelligence in India – 2020)[8]

2.1 AI-based Speech Technologies

For humans, speech is the most natural form of communication, and at present, machines have also started interacting with us through speech. Since the 1930s, researchers have made efforts to build systems that can recognise, understand, and produce speech. These efforts gained massive momentum with the growing availability of large amounts of digital data, an increase in computational power, and the advent of efficient deep learning models and algorithms.

When it comes to the domain of speech technology in Artificial Intelligence, there are two main applications or rather technologies:

ASR (Automatic Speech Recognition) and TTS (Text to Speech). A typical voice-based conversational AI usually consists of the following modules:

- Automated Speech Recognition (ASR)
- Natural Language Processing (NLP), which includes Natural Language Understanding (NLU) and Natural Language Generation (NLG).
- Text to Speech (TTS)

NLP, which helps computers understand, interpret, and manipulate human language, is a process that acts as an enabler to both ASR and TTS.

Around 90 percent of the Indian population in the 16 to 64 years age group owns a smartphone. As of January 2020, about 50 percent of the Indian people are mobile internet users.

2.1.1 Automatic Speech Recognition (ASR)

Automatic Speech Recognition or ASR allows human beings to speak with a computer interface in a way that resembles normal human conversation. ASR can be defined as the decoding of linguistic information from speech signals with a machine's help. Information about the speaker and language is extracted through ASR. In an ASR system, a user speaks out some words recorded through a microphone by a machine. The quality of the audio depends on the recording device or microphone used.

After the voice is recorded, signal processing is performed wherein the machine computes speech features that will help identify and characterise speech sounds present in sound waves. This characterization process is also known as feature extraction. The most-used method of transforming a sound wave is Fast Fourier transformation and then creating a spectrogram for feature extraction. Transformation reduces the dimensionality of sound data so that individual words and, ultimately, sentences could be predicted out of it. This way, the machine can transcribe what has been said and effectively convert speech into text.

2.1.2 Text-to-Speech (TTS)

TTS, also known as "read aloud" technology, reads out digital text through a computer-generated voice. A computer/machine used for this purpose is called a speech synthesizer.

The TTS process initially involves pre-processing of the text not to cause any errors at later stages when the machine reads it out. Things like dates, abbreviations, times, numbers, acronyms, and special characters (like currency symbols) need to be turned into words.

This step is not as easy as it seems. For example, take the number 1985. Humans can decipher whether this is a number, a year, time, or padlock

Figure 4: Smart Speakers (Left: Google Nest Mini, Right: Amazon Echo)

formation by their sense of what is written and the context behind it. A machine does not have that sense in-built; hence it uses statistical probability techniques (like Hidden Markov Models) or neural networks to arrive at the most likely pronunciation.

If the word year has been used earlier in the sentence, it is most likely a year and will be pronounced accordingly. Homographs are another issue that pre-processing addresses. Homographs are words pronounced in different ways to convey different meanings, e.g., 'read.'

Finally, the text is converted into a list of phonemes (sequence of sounds given in a dictionary alongside a word). These phonemes have corresponding audio pronunciations, either recorded in the human voice (concatenative approach) or simulated by generating fundamental sound frequencies (formant approach).

A third method, called articulatory, is still being explored where computers speak by modeling the human vocal apparatus. It is the most challenging approach out of the three and has not yet been used in today's common speech systems. This way, the machine can read out written text and effectively transform the text into speech.

2.2. Speech System for Indian Languages

Nowadays, speech recognition technologies have been available popularly in smart speakers such as Amazon Echo or Google Home. These platforms support communication in Indian languages to a limited extent only. While Amazon Alexa supports Hindi among seven other international languages, Google Home supports 13 languages, including Hindi, as the only Indian language[24]. Microsoft supports Indian English, Hindi, Tamil, Telugu, Gujarati, and Marathi for its ASR systems.

When it comes to TTS, Amazon Polly supports various international languages but none of the Indian languages. Limited coverage of Indian languages is also the case with Microsoft Azure, which supports 45 languages (including Hindi, Tamil, and Telugu), but other Indian language adaptation has not yet been addressed[12].

Although large parts of the voice AI market in India are captured by tech giants such as Google, Amazon, Microsoft, etc., startups are also trying to venture into it and leverage the research and development made in the Speech and NLP field so far.

In a country like India, which has 22 official languages (scheduled) and more than 720 dialects[13], local languages like Assamese, Punjabi, Dogri, etc., remain unexplored developing modern-day NLP and speech systems. In recent years, NLP has been mostly focused on data-driven approaches rather than the earlier rule-based systems. Therefore, today's NLP (and speech) based systems demand large amounts of data which is well annotated and transcribed to train models using advanced deep learning techniques to provide higher accuracy.

Dependency on AI applications has increased with new voice-interactive devices like smartphones, smartwatches, and smart home appliances. The major bottleneck in the development of Indian language technologies has been the lack of linguistic resources to contribute to these devices' smartness and make them work for Indic languages. The lack of digital text data in many Indian languages complicates innovation in NLP in India.

The major bottleneck in the development of Indian language technologies has been the lack of linguistic resources to contribute to these devices' smartness and make them work for Indic languages.

Research: Objectives & Approach

3.

3.1 Research Objectives

'FAIR Forward – Artificial Intelligence for All' strives for a more open, inclusive, and sustainable approach to AI at the international level. The focus is on removing entry barriers to AI by providing open access to speech datasets in Indian languages.

The primary objective of this research is to identify and evaluate existing (open) speech datasets available in Indian languages. The research activities include:

- Mapping existing open voice datasets in Indian languages. This exercise aims at identifying the existing open voice datasets for Indian languages through desk research, professional networks, qualitative interviews, and related methods.
- Identifying and evaluating the datasets and methods been chosen to collect them.

  Here the identified datasets are further examined to retrieve information about their quantity, quality, and the approaches that have been adopted to generate them. The collection mechanisms are studied in detail to find the reasons behind their selection and the advantages/disadvantages of using them.
- Outlining the possible recommendations for future sustainable voice data collection.

  Based on the above observations, recommendations are being made to address the unavailability of quality Indian language speech datasets.

3.2 Research Approach

A phase-wise approach was taken to address the objectives laid down for this research work. The two phases executed were as follows:

Phase 1: Online Secondary Research

The first phase of the research involves performing online secondary research of the information regarding voice data in Indian regional languages available in the public domain. This includes:

- Searching for existing voice datasets that are available in Indian languages.
- Retrieving more information about such datasets concerning their volume, quality, collection mechanism, etc.
- Finding out prominent researchers, institutes, and business enterprises who have worked on or are working in the field of speech technology using Indian language voice datasets.

Here, the focus has been to explore and gather information about voice datasets that are open-source or could potentially become open-source.

Various data sources related to TTS/ASR were screened, evaluated, and listed down: Around 40 sources were identified, out of which 29 sources were found to be significant in terms of volume, number of speakers, and languages covered.

A detailed list of these 29 significant data sources, along with their access links, can be found in Appendix 4.

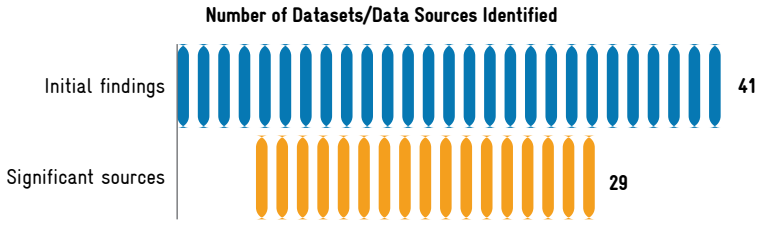


Figure 5: Number of Indian speech datasets listed during secondary research

FAIR Forward – Artificial Intelligence for All' strives for a more open, inclusive, and sustainable approach to AI at the international level.

Phase 2: Qualitative Interviews

This includes qualitative interviews with researchers, institutes, and enterprises that were identified during Phase 1.

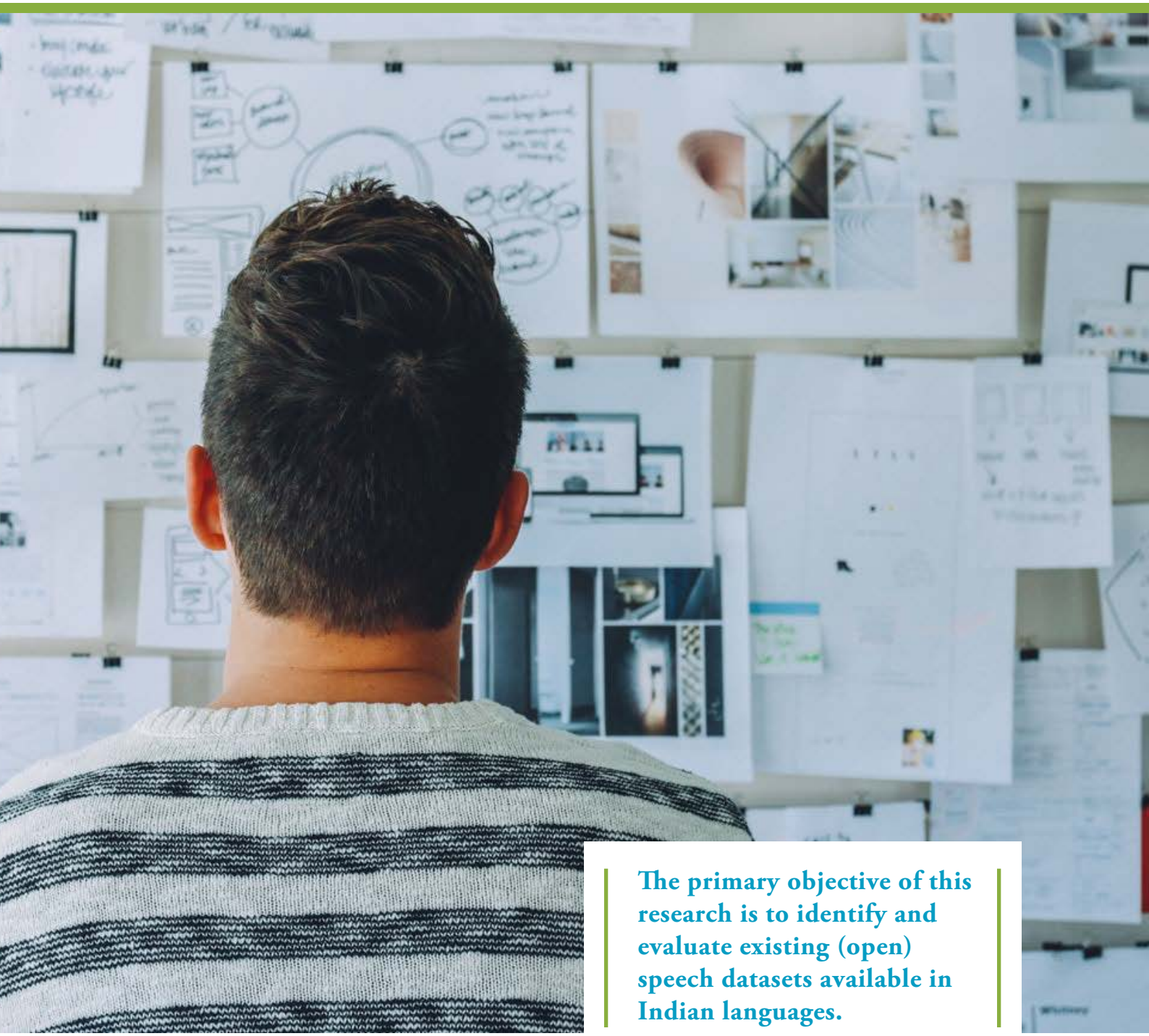
An email request was sent to various prominent researchers and industry experts active in the voice AI domain for Indian languages.

All those who kindly agreed to contribute to the present study were individually interviewed over an online session. They were asked a series of qualitative and quantitative questions relating to the current state of voice AI in Indian languages and the way forward.

Around 15 experts were interviewed, out of which eight were researchers/professors from renowned academic and research institutions. Five interviewees were key stakeholders/ founders of startups/ business enterprises/NPOs (Non-Profit Organizations), while the remaining two were heads of relevant departments in government organizations.

Finally, based on the findings from Phase 1 and 2, recommendations were collated for future creation and sharing of voice datasets.

Figure 6: The Research Process Flowchart



The primary objective of this research is to identify and evaluate existing (open) speech datasets available in Indian languages.

4. Understanding Voice Data

As observed with other AI technologies, speech technology requires data at its core to be functional. These datasets must be available in individual formats specific to ASR/TTS systems. For building an ASR system, a speech dataset/corpus must include:

a) **Audio files:** Contain spoken words or sentences in a noisy or noise-free environment.

b) **Transcripts:** Transcripts are the textual data for the speech in audio files.

c) **Annotations:** The transcribed data must be annotated. Annotating is the process of labeling speech information to identify linguistic details like names, objects, verbs, etc. This POS (Part of Speech) tagging process of speech data helps the machine learning algorithm learn more effectively.

d) **Metadata:** Information related to the speaker's demographics and the language spoken (including the dialect) also needs to be captured. This is because, for an ASR system to work efficiently, it must accurately decipher what is spoken. The accuracy can only be achieved if the model is well trained according to the end-user voice characteristics based on their gender, age, etc., and the dialect of a particular language.

Figure 7 gives an idea about the top 10 Indian languages based on volume (in hours) of ASR data available (Appendix 3) for those languages.

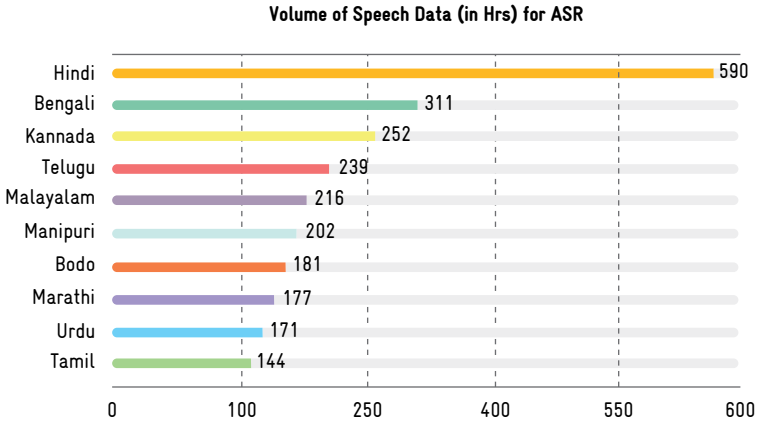


Figure 7: Top 10 Indian languages based on their data volume (in hours) for ASR (Appendix 3)

For building a TTS system, a speech dataset/corpus must include:

a) **Audio files:** Contains spoken words or sentences. (preferably in a noise-free or studio environment)

b) **Pronunciation Lexicon:** This is a mapping of words to their corresponding pronunciation. written in a particular convention (like IPA - International Phonetic Alphabet)

c) **Phonological definitions:** This defines all the possible phonemes (sounds) of a language and its corresponding definitions. For example, it lists all possible vowels and consonants and indicates properties like articulation, nasality, etc.

ASR models require rigorous training to develop systems that can understand speech from multiple users with different accents in various situations (environments).

A TTS system needs to synthesize speech based on the text provided to it, for which its must be trained on voice data generated by a single speaker.

Therefore, building an ASR system requires a larger volume of speech data from various speakers in different environments (noisy/noise-free) compared to a TTS system, which needs to be trained on speech data from a single speaker. TTS systems usually operate on data collected in a noise-free environment.

Building an ASR system requires a larger volume of speech data from various speakers in different environments (noisy/noise-free) compared to a TTS system, which would work even when trained on speech data from a single speaker.

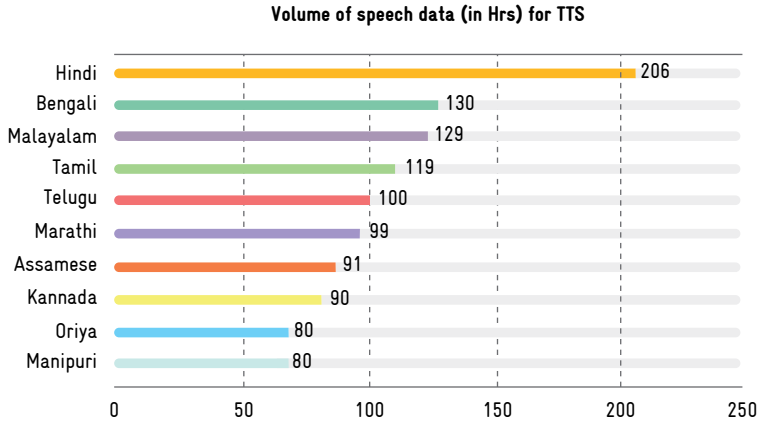


Figure 8: Top 10 Indian languages based on their data volume (in hours) for TTS (Appendix 3)

Hence, separate data collection exercises must be initiated for ASR and TTS systems, respectively, as their data are not suitable to be used interchangeably.

Figure 8 gives an idea about the top 10 Indian languages based on volume (in hours) of TTS data available (Appendix 3) for those languages.

4.1 Availability of Data

In an endeavor to build, deploy, and scale voice-based AI systems, the availability of good quality and enough data is a prerequisite. The available datasets are presently hosted on different online platforms

4.1.1 Data Access/Procurement Procedures

Every data hosting platform has specific sets of procedures following which a dataset can be downloaded. The typical steps a user must follow to access the data on various platforms are:

- User simply clicks on the download link to download the corpus from the webpage (e.g., OpenSLR corpus).
- Users must register themselves on the website/platform to download the dataset (e.g., Mozilla Common Voice).
- If the data is not available for free, then the user post-registration must pay for the data to access it (e.g., various datasets on ELRA, LDC).
- In case the dataset is available (free) only for research and academic purposes, then the user must fill in additional details about himself/herself and affiliation with a project (or institute). Post verification from the data host, the dataset is made available to the user (e.g., LDCIL datasets).

The following are examples of datasets that are easily accessible and involve minimum efforts from the user side to download them:

1. Open SLR: Single click download of the dataset without providing any user information. (http://www.openslr.org/resources.php)
2. LibriVox Audiobooks: Single click download of the dataset without providing any user information. (https://librivox.org/search)
3. Festvox (CMU INDIC): Single click download of the dataset without providing any user information. (http://festvox.org/h2r_indic/)

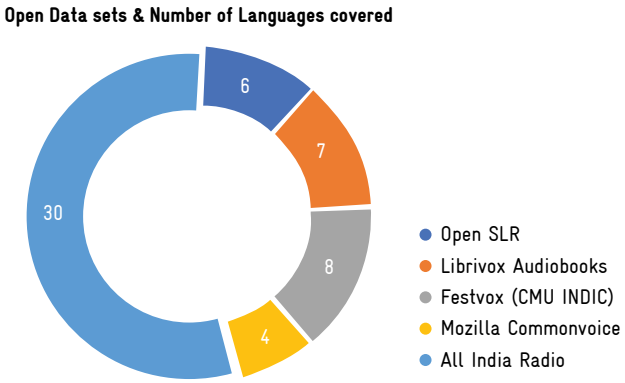


Figure 9: Easily accessible datasets and the number of Indian languages they cover

4. Mozilla Common Voice dataset: The user must provide an email address before downloading the available dataset. (https://commonvoice.mozilla.org/en/datasets)

5. All India Radio (AIR): Menu and selectiondriven download process to retrieve audio and transcriptions of news bulletins (e.g., selecting Date, Time, Region, Language, etc.). However, the number of speakers is limited, thus a challenge. (http://newsonair.com/RNUNSD-Audio-Archive-Search.aspx)

The existing speech datasets available with different research groups / academic institutes are limited in volume and driven by a specific use case. This can be attributed to the fact that such datasets were collected as a part of projects with specific objectives, and data collection was never the primary goal of these projects. A dataset captured during a project aimed at building home assistance ASR systems cannot build a general-purpose ASR system.

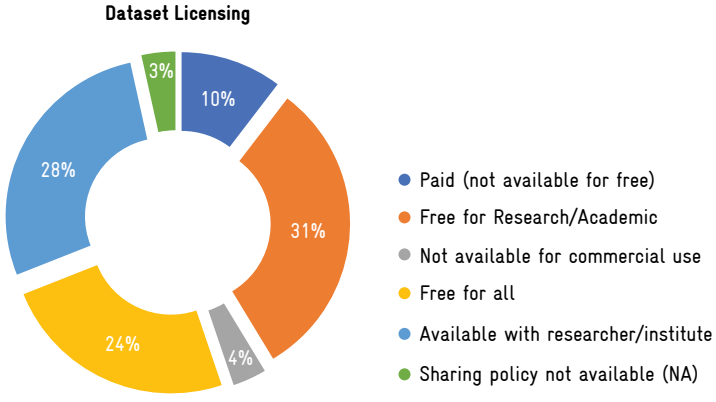


Figure 10: Dataset licensing/availability (Appendix 3)

4.1.2 Dataset Licensing

There is no common platform or interface through which all (or a majority) of the existing datasets from various sources could be accessed. However, NPOs like the EkStep Foundation have been working on filling this gap. EkStep is working to generate thousands of hours of speech data in Indian regional languages and publish it as open-source. The methodology followed by them is to web scrape audio files from various sources and use AI techniques to cleanse it. Later these audio files are transcribed using amazon cloud services. Microsoft Research has also undertaken the task of collecting speech data in Indian languages with an intent to make them open-source. It has already collected a few hundred hours of speech data, expected to be made public soon.

The speech datasets for Indian languages owned by large technology companies like Amazon and Google are generally not available publicly. Most of the freely available datasets were generated by a researcher or an institute during some project/research. From the datasets shortlisted through this research, it can be inferred that approximately 31 percent of those datasets are available free of cost for non-commercial purposes (and paid for commercial usage). 24 percent of the datasets are freely available for all purposes, while 28 percent of the datasets are available only with related researchers and institutes, and there is no clear distribution policy associated with them at present. (Figure 10).

Examples of datasets free for research/academic purposes are the ones available under LDCIL (Linguistic Data Consortium for Indian Languages). Here the datasets are available for both commercial and non-commercial users. However, for commercial users, they are chargeable.

Out of the 29 identified resources (Appendix 3), about 31 percent are available on government/NPO platforms, 55 percent are from academic/research institutes, and the remaining 13 percent are from business enterprises.

A researcher or institution is often governed by the funding organizations' policies (government or private) that provide them with financial support to complete a project. In many cases, the researcher or institute is uncertain of their authority to make independent decisions about the dataset's licensing and distribution elsewhere.

This ambiguity is the present situation with many datasets that are available with institutes or researchers. TDIL (Technology Development for Indian Languages) is a scheme by MeitY (Ministry of Electronics & Information Technology), Government of India, under which many speech-related projects have been funded. These projects have been executed by prominent institutions like IIT (Indian Institute of Technology) Madras, IIT Guwahati, IIIT (International Institute of Information Technology) Hyderabad, among many others. However, there are also scenarios where data generation activity is conducted independently. In such cases, the researcher or institute may have the liberty to decide on making those datasets openly accessible. Due to the lack of any widespread initiative, such datasets are hardly available in the public domain. A researcher generally does not have any issues making the datasets open to the public, provided the due credit and acknowledgment be given to their contribution.

One such example is independent research done at BAMU (Dr. Babasaheb Ambedkar Marathwada University) that has resulted in several hours of speech data in Marathi language for the agricultural domain, focusing on crop diseases and their remedies. This dataset is not yet available in public, but the researchers are keen to open it up for public use.

There are cases where startups have also taken up the painful exercise to collect data to fulfill their requirements. For example, Gram Vaani, a tech-startup incubated at IIT Delhi, was set up to empower communities using relevant technologies. Graam Vaani has so far collected 130 hours of speech data in Hindi. This data has been made available on the ELRA (European Language Resources Association) platform, which can be used free of cost for research/academia purposes and is paid for commercial purposes. The revenue generated from the distribution of these datasets helps sustain the startup.

A comprehensive view of the speech datasets available in Indian languages is available in Appendix 3. Out of the listed datasets, the most prominent platforms with hosted voice data (based on volume/languages/number of speakers) are shown in Table 2:

- **LDCIL** (Linguistic Data Consortium for Indian Languages) - set up under MeitY, Government of India, in collaboration with CIIL (Central Institute for Indian Languages) is a data hosting platform similar to LDC, but for Indian languages. (https://data.ldcil.org/)
- **NPLT** (National Platform for Language Technology) - set up under MeitY, Government of India, is a platform for academia, researcher, and industry to provide Indian language data, tools, and related web services. (https://nplt.in/demo/)
- **PM Mann Ki Baat -** Prime Minister's (PM) Mann Ki Baat is the monthly address of Indian PM Narendra Modi to its citizens through recorded audio. This includes conversations with other callers as well. The website hosts multiple episodes with audio files and corresponding transcripts.

  (https://www.pmindia.gov.in/en/mann-ki-baat/)

4.2 Data Collection Mechanisms

There is no specific go-to mechanism when it comes to speech data collection in India. As long as the data collected hasA the desired quality and serves the intended purpose, any cost-effective and time-efficient mechanism will work. The data collection methods can be categorised based on factors such as speech type, the recording environment, recording devices, etc.

Table 2: Prominent data hosting platforms based on volume of data, number of languages and number of speakers

Hosting Platform	Volume of dataset (hours per language)	Indian Languages covered	Number of Speakers (per language)
LDCIL	119	13	385
NPLT	36	14	1500 (ASR), 2 (TTS)
TDIL	-	5	1119
All India Radio	>1000	>30	-
PM Mann Ki Baat	>35	11	1-2 per episode

- **TDIL** (Technology Development for Indian Languages) is a program initiated by MeitY to create and access multilingual knowledge resources and using them to develop innovative voice AI-based products and services for common users. (https://tdil.meity.gov.in/Default.aspx)
- **All India Radio (AIR)** - AlR is the national public radio broadcaster of India and a Prasar Bharati division. Their website hosts hundreds of audio files and corresponding transcripts for news bulletins and programs broadcasted over the radio. (http://newsonair.com/)

4.2.1 Speech Types

The following are some of the common types of speeches that are generally recorded:

a) Read speech (guided speech): A speaker reads out the transcripts provided to him/her, also called guided speech.

b) Extempore: A speaker is given a topic (or a question) and is supposed to speak about it impromptu.

c) Conversational/Debates: These are recordings of a natural conversation between multiple (mostly two) speakers.

d) Lecture: Lectures on various topics are recorded as delivered in educational institutes.

Among the speech-types mentioned above, guided speeches have been the preferred method for capturing speech data for Indian languages. This is because, in India, languages (and their accent/dialect) change every few hundred miles, and speech in Indian languages is highly colloquial. Thus, having a transcript ready and verifying the spoken words with the text at hand makes the process time-efficient and ensures quality. In addition to this, there is assurance that guided (or read) speeches would be phonetically rich/ balanced and would cover all the required domain-specific vocabulary. Microsoft Research has followed this methodology to collect speech datasets for Indian languages.

The transcript and the speaker selection are made based on the type of speech system to be developed (e.g., ASR or TTS) and its overall objective. For example, if an ASR system needs to be developed for farmers in Tamil Nadu, the transcript must cover all subjects related to agriculture and its issues. The speaker should also be a farmer from a village in Tamil Nadu whose native language is Tamil. At the same time, the same exercise is conducted with non-native speakers as well to cover various accents of the same language. The transcripts have to be carefully drafted by experts to contain phonetically balanced sentences and cover all the domain-specific vocabulary.

4.2.2 Data Collection Process

Over time, there has been a gradual shift in the approach adopted for collecting speech data. In recent years, the focus has been laid on automating the collection activity and reducing manual efforts.

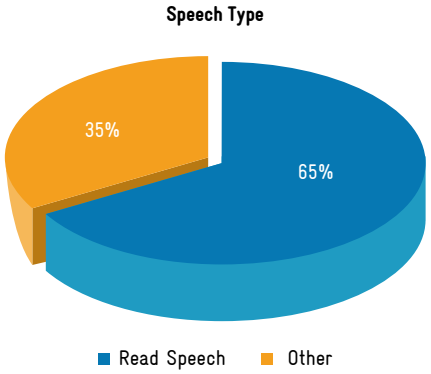
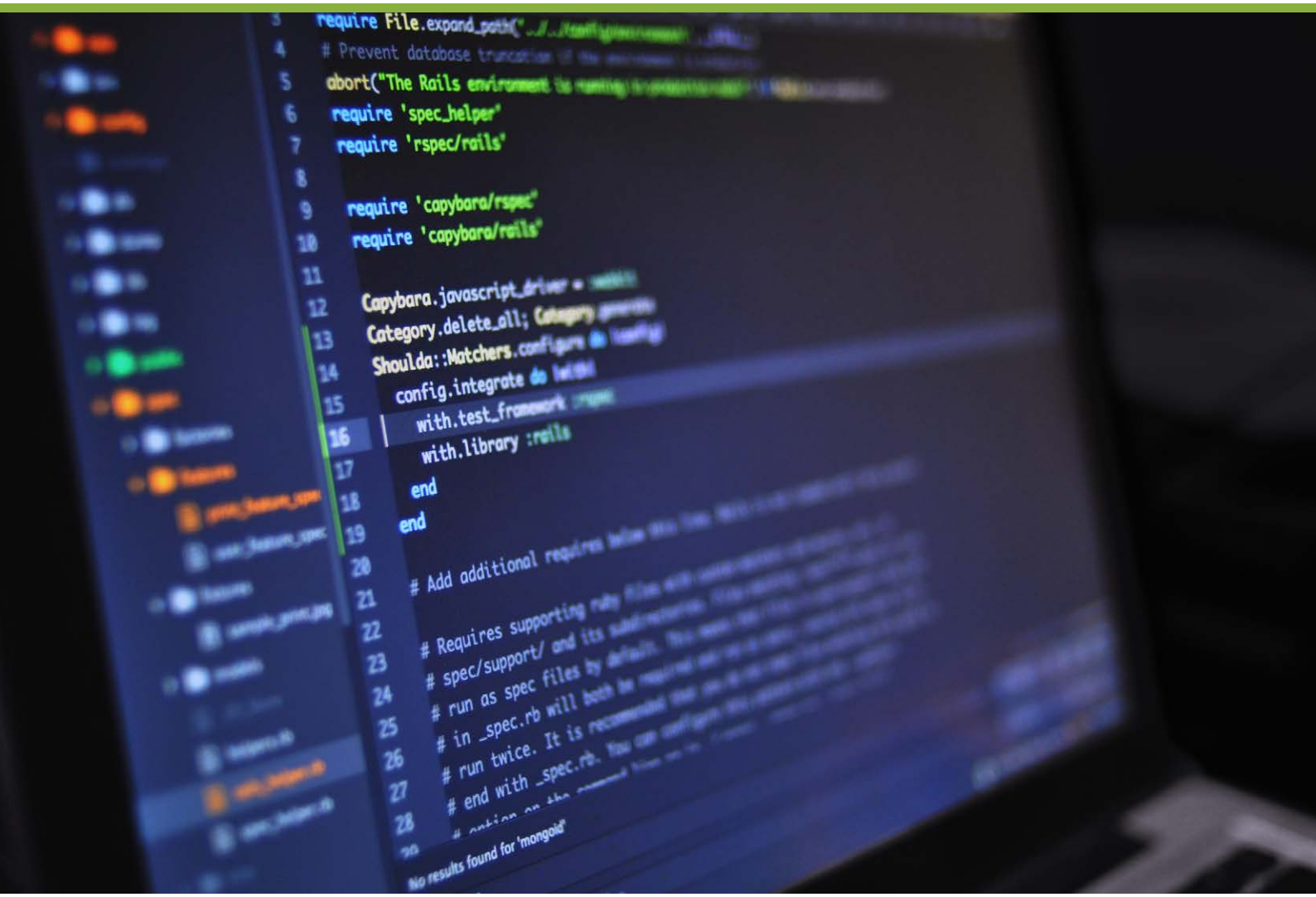


Figure 11: Percentage of datasets from Appendix 3 categorized based on Speech Type

4.2.2.1 On-Site Data Collection

The first step in the conventional process being followed for voice data collection in Indian languages is to identify relevant speakers. Next, a team visits the location (depending upon the language to be captured) with the necessary equipment. They rent/set up an infrastructure to get speakers' voices recorded in an appropriate environment.

Organizations like CDAC (Centre for Development of Advanced Computing), LDCIL, and many other autonomous academic institutes like IITs have been following this conventional approach.



Apart from having the advantage of capturing local dialects, this approach also has drawbacks in terms of excessive time and resources it consumes. With affordability and a higher reach of smartphones and the internet, organizations are now adopting smarter methods wherein the researcher does not have to visit various locations to record the speech data manually. Gram Vaani has been successfully collecting relevant data from local people through an IVR mechanism over mobile/telephone networks.

4.2.2.2 Online Data Collection

Methods like crowdsourcing have gained recent attraction in which an application or a website hosts a large quantity of textual data, and anyone willing to contribute can read out these texts through a microphone and get their speech recorded. The verification (or data cleansing) process can also be crowdsourced by making the recordings publicly available. People can listen and verify the recordings with their corresponding transcripts. This saves significant time and resources. Mozilla Common Voice is one such crowdsourcing platform where the data is recorded and processed by the people themselves.

Vakyansh (EkStep Foundation) is another volunteer-based crowdsourcing platform.

However, unlike Mozilla Common Voice, Vakyansh captures speaker information (metadata) in the form of age, gender, and native language. Microsoft Research under project Karya has also been collecting crowdsourced speech data. However, the problem with crowdsourcing is that there are no physical means to verify the speaker's demographics. Also, if there is no incentive associated, people would not be motivated to give their voice for the speech dataset. Karya has tried to address this issue by recruiting appropriate speakers on fixed wages to encourage them to donate their voices.

TDIL and CDAC have recently adopted application-based collection techniques where transcripts to be read are provided to speakers over mobile applications.

data from their daily news and debate shows. However, transcribing a pre-recorded speech can be tedious for Indian languages, which are numerous and have multiple dialects. Unlike the English language, Indian languages still lack efficient AI tools to automate the transcription activity.

Government bodies like TDIL and CIIL have mostly relied on autonomous institutes like (IIT Madras, IIIT Hyderabad) or other government bodies like CDAC to collect speech data. LDCIL has been setup as a common platform for hosting language datasets and resources, but the data generated so far remains inadequate to develop a general purpose ASR system.

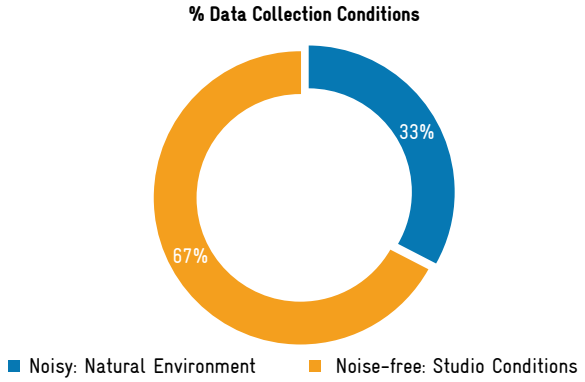


Figure 12: Percentage of datasets from Appendix 3 categorized based on Recording Environment

There are methodologies where the initial focus is to collect the speech/voice samples. The transcription of these files happens at a later stage. EkStep Foundation has collected thousands of hours of voice data through web scraping. They have an automated system deployed to pre-process and clean the data gathered. These audio files are then transcribed through Amazon Cloud Services. As there are no manual quality checks in place, the quality of such datasets is yet to be ascertained.

CIIL/LDCIL is also planning to partner with news channels like DoorDarshan (DD) to access their voice data. Other organizations are also looking forward to collaborations with media houses and news channels to get speech

Large enterprises like Amazon, Google, etc., hire MSMEs/startups to collect data for themselves. These datasets may never come into the open domain. Apart from prominent organizations and government bodies, small and medium-sized enterprises (SMEs) and tech startups do not have enough resources and funds to collect the data for themselves. Therefore, they provide AI and NLP solutions based on the datasets provided to them by their clients/partners like media houses, educational institutes, or government organizations.

As an example, Devnagri, an AI-based startup has thousands of hours of voice data in English provided by their clients/partners. The data pertains to the educational domain and has been transcribed and translated into Indian languages using AI and human intervention.

4.2.3 Recording Environment

The recording environment is another crucial aspect of any speech data collection mechanism. Speech data is collected in two types of environments: Noisy and Noise-free. The environment in which the data is collected determines the applicability of the speech system being developed.

For example, to develop an ASR system that can be used for navigation while driving, the system needs to be trained on speech data captured in a similar noisy environment (road/highway).

When it comes to the available speech datasets in Indian languages, the most common setting in which the data has been recorded is a closed-room environment (e.g., office, home, etc.) or a studio environment. Studio recordings are preferable for building TTS systems in general.

Since each data collection method has its advantages and disadvantages depending upon the overall objective of the task at hand, the voice data collection mechanisms discussed are summarised in table 3, along with their pros and cons.

4.3 Quality of Voice data

Based on the interviews with experts, the quality of any speech dataset is measured by its usefulness to meet the speech system's objectives, and the type of speech systems (ASR/TTS) being developed.

The datasets collected by IIT Madras are majorly for speech synthesis (i.e., TTS), whereas those collected by IIIT Hyderabad are mostly for ASR systems. These two sets of data have different quality, and they are not suitable to be used interchangeably. The quality is also influenced by factors such as the recording device, recording environment, types of speakers, etc. In the following sections, each of these factors has been discussed in detail.

Table 3: Voice data collection mechanisms with their pros and cons

	Pros	Cons
Speech types		
Guided Speeches	Reliable, Suited to Indian languages	Need to hire expert to prepare transcripts
Extempore	Captures vocabulary and ways of speaking that are more relevant for real world applications	Low reliability due to lack of transcripts and high dialectical variations
Conversational	Captures vocabulary and ways of speaking that are more relevant for real world applications	Low reliability due to lack of transcripts and high dialectical variations
Lectures	Suited for domain specific ASR/TTS	Not suitable for general purpose ASR/TTS
Recording environment		
Noisy	Helps in building robust speech systems (ASR)	Denoising (removing noise) required for unexpected noises, not suitable for TTS systems
Noise-free	Easy for sound processing	Not suited for real world application
Collection process		
On site collection	Good quality	Time consuming
Online collection	Less arduous	Quality not assured

4.3.1 Quality Based On Recording Device Used

Audio quality depends upon the device used to record the sounds. Nowadays, there are minimal challenges in procuring devices needed to capture and generate quality sounds. The microphones have a frequency response in the desired frequency range. Also, good A2D (analog to digital) converters are available, and one can get a high number of bits per sample through them.

Any recording done through a mobile phone or laptop would most likely be of mono (channel) quality, while recording from an advanced microphone, or a studio setup will result in stereo (or multi-channel) quality audio data. Different quality levels might be needed for various purposes. For building a mobile phone ASR system (like IVR), mono audio will suffice. However, it is good to have studio recordings to understand the basic patterns of the speech sounds. Multi-channel (stereo) recordings can also help cancel noise (denoising).

Hence, to record, store, and process speech data, there is no limit to the type of advanced devices that can be used. However, at present, mobile phones, laptops, and headphones remain the most cost-effective, readily available, and widely used recording devices.

4.3.2 Quality Based On The Recording Environment

The recording environment's selection should be based on the surroundings in which the speech system will be deployed. Here SNR (signal to noise ratio) becomes a susceptible factor. Also, speech data for TTS systems must be noise-free. Therefore, it should be recorded in a studio-like environment.

At present, most of the available voice datasets in Indian languages have been captured in either a closed-room environment (e.g., office, home, etc.) or a studio environment. Thus, such datasets may have limited success rates if used for real-life applications where user interaction happens in a natural setting.

The datasets collected through IVR systems by the Gram Vaani team are noisy as they have been captured from people speaking in an open natural environment. Apart from the excessively noisy ones, these datasets help build ASRs that are robust and accustomed to working in real-life situations.

Datasets on LDCIL portal have been captured in natural environment and are useful for building ASR systems.

4.3.3 Quality Based On The Method Of Transcription

It is found that the speech transcripts presently available are formed with the intent to have sentences that are phonetically balanced and cover frequently used words (like names, places, dates) of an Indian language. In addition to these, the domain-specific vocabulary is also included in transcriptions for the speech system to work as per the situational requirements (e.g., banking, education, agriculture, etc.).

The transcriptions must also be well-annotated. It is advisable to involve domain and linguistic experts to prepare acceptable quality transcripts.

It is hard to find good quality, well-annotated, and transcribed speech datasets for all the significant domains like finance, education, healthcare, etc., in Indian languages. However, agriculture is one domain where some progress has been made as a part of the Mandi project under TDIL. Mandi Project was undertaken to help farmers stay updated with the latest price for agricultural commodities and weather reports through an IVR system.

4.3.4 Quality Based On The Variety Of Speakers

Identifying the demographic information about a speaker is also essential to ensure the coverage of different types in accents/dialects for a language. The accent, choices of words, and sentence formation of a middle-aged, highly educated urban class individual would differ from an aged, not formally educated rural-based individual. Even for the same language, some dialects are so different that each dialect must be treated as a separate language altogether.

Gender and age group are the two most common demographic characteristics captured in currently available datasets, where metadata has been recorded.

4.3.5 Quality Based On Domain Coverage

Apart from big enterprises like Google and Amazon, there has not been any serious effort made by any researcher or institute to build a large-scale general-purpose speech system for Indian languages.

The volume of the datasets available with various researchers and institutes are in accordance with the requirements of their limited research work and specific projects. Hence, it is not certain that those existing volumes of datasets will meet a purpose different from the project under which they were collected.

For example, the Indian Institute of Science (IISc) has conducted a research project on building an ASR system for whispers. Here, the training dataset consists of audio files of whispers and corresponding normal voice. Such datasets are explicitly built for whisper ASR systems and may only enhance existing ASR systems' accuracy.

Gram Vaani has followed the bottom-up model of information flow, wherein a local user discusses and raises queries related to various domains like agriculture, health, sanitation, etc., over an IVR system. This query is then transcribed through an automated transcription service of Amazon using custom vocabulary.

However, automated transcription might not give highly accurate results for Indian languages. The query is then matched with the existing database of questions and answers (Q&A). If the query has a context similar to any question that exists in the database, corresponding pre-recorded responses are returned to the user. Therefore, the domain is limited to the topics covered in the existing database, and the system will fail to produce an automated response (answer) if an out of context query is asked.

Similarly, the data available on the TDIL portal is related to the agricultural domain (as a part of the Mandi project), whereas LDCIL claims to have more general-purpose datasets. Microsoft Research also claims to have collected a few hundred hours of general-purpose speech data for the Hindi language.

The NLTM/ Bahubashak pilot project aims to build speech-to-speech translation systems for NPTEL (National Programme on Technology Enhanced Learning) lectures.

4.4 Volume of Voice Data

The volume of a speech dataset is generally measured by the number of hours of audio recordings. The number of speakers and languages covered is also a good indicator of the expanse of a dataset.

Due attention needs to be paid to the following scenarios to judge what volume of speech dataset will be enough to build an efficient speech system:

The type of speech system being built:

An ASR system would generally require hundreds to thousands of hours of data, while a TTS system might work even with 25-50 hours of data although from single speaker.

The language being covered:

The variety of phonemes and the number of derived words in every language is different. For example, in Tamil, the root word 'wa' (English: come) might have 500 variations (derived words), while 'come' in English has only limited variations like 'coming' and 'came'.

The number of dialects in a language:

Each Indian language has a different number of dialects, and to build a speech system for an Indian language, all the dialects must be covered. A language with 20 dialects requires more volume of training data than a language with only three dialects.

The domain of the speech dataset:

Adequacy of data is a derivative of the domain chosen. A vocabulary that covers crop diseases might be smaller in volume than the vocabulary for human illnesses. Therefore, compared to the agricultural domain, a larger volume of speech data would be needed for the human healthcare domain.

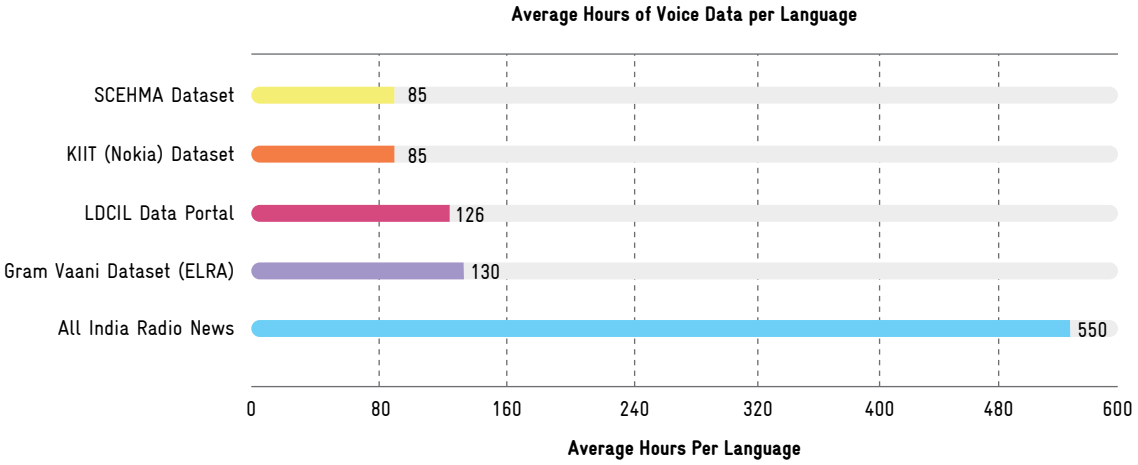


Figure 13: Top 5 datasets based on volume (hours) per language (Appendix 3)

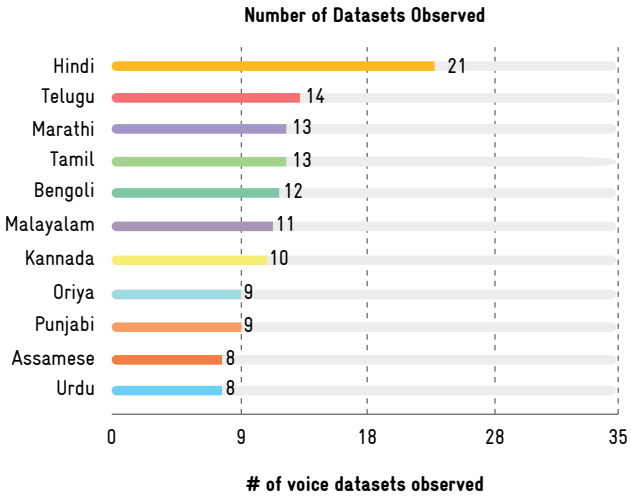


Figure 14: Top 11 languages based on the number of datasets they are captured in (Appendix 3)

NLP techniques and algorithms:

Some NLP (and speech) algorithms use small volumes of data to learn, whereas others require large amounts of training data. The open datasets available are limited in volume because they were generated for a specific purpose as part of a project whose primary focus was never data collection.

4.4.1 Volume Based on Hours of Speech Data

There is an acute need for a long-term sustainable project to collect large volumes of good quality speech data in Indian languages. Based on the number of hours (per language average) of voice data, the four noteworthy speech datasets for Indian languages are shown in figure 13.

LDCIL is one platform that was developed by the government to serve as an LDC replica for Indian languages. LDCIL has speech datasets for 19 Indian languages with an average volume of 126 hours per language.

The NLTM /Bahubashak pilot project aims to capture about 3000 hours of Hindi, 2000 hours of Indian English, and 1000 hours of Tamil speech data.

The Vakyansh (EkStep Foundation) crowdsourcing platform has gathered more than 33 hours of speech data from more than 370 speakers. The platform aims to collect 10000 hours of speech data in Indian languages.

Through Karya (crowdsourcing platform of Microsoft Research), around 500 hours of speech data for the Hindi language has been captured. Microsoft Research in collaboration with Navanatech aims to collect 2000-3000 hours of speech dataset for Oriya language (with four dialects) in healthcare, finance, and agricultural domains.

parts of the country. Significant efforts are yet to be made to capture all the dialectical variation of the 22 official languages.

The majority of the existing datasets cover the Hindi language. Telugu and Marathi are the other two predominant languages covered.

Since most of the projects and research related to speech technology (for Indian languages) have been use-case (or domain) specific, the datasets captured are consequently very limited with respect to the languages and dialects covered.

On the contrary, there are platforms where one can find speech datasets in numerous Indian languages, even though the platforms and their resources were never intended to be used for building ASR/TTS systems. Examples of such platforms are PM Mann Ki Baat or All India Radio news archives. The sanctity of datasets from such platforms needs to be verified before they could be put to actual use.

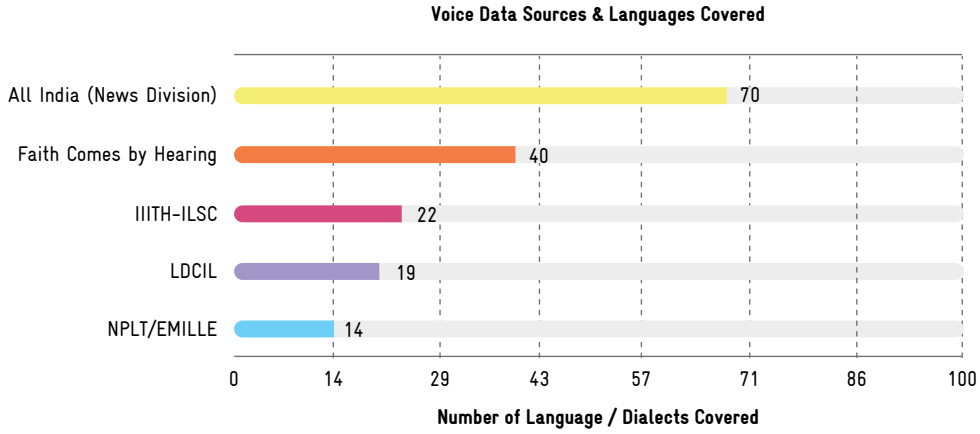


Figure 15: Top 5 datasets based on number of languages covered (Appendix 3)

4.4.2 Volume Based on Number of Languages Covered

Most of the speech datasets available for Indian languages cover a subset of the 22 official languages.

Almost no efforts have been made to capture any of the non-scheduled languages. However, each of the 22 scheduled languages also has multiple dialects based on how they are spoken in different

CIIL/LDCIL has captured data on 18 of the scheduled Indian languages, even though their volume might not be sufficient for a robust general-purpose voice AI system. 'Sanskrit' and 'Sindhi' are the two languages where no progress has been made so far to capture speech data.

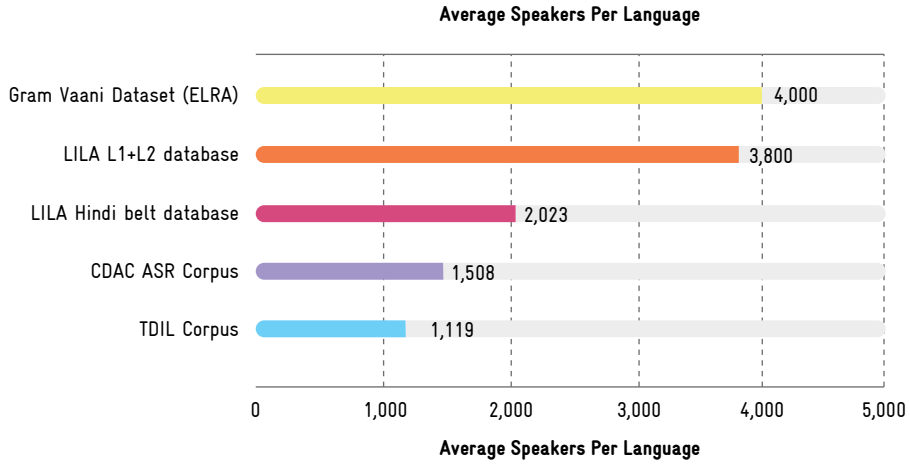


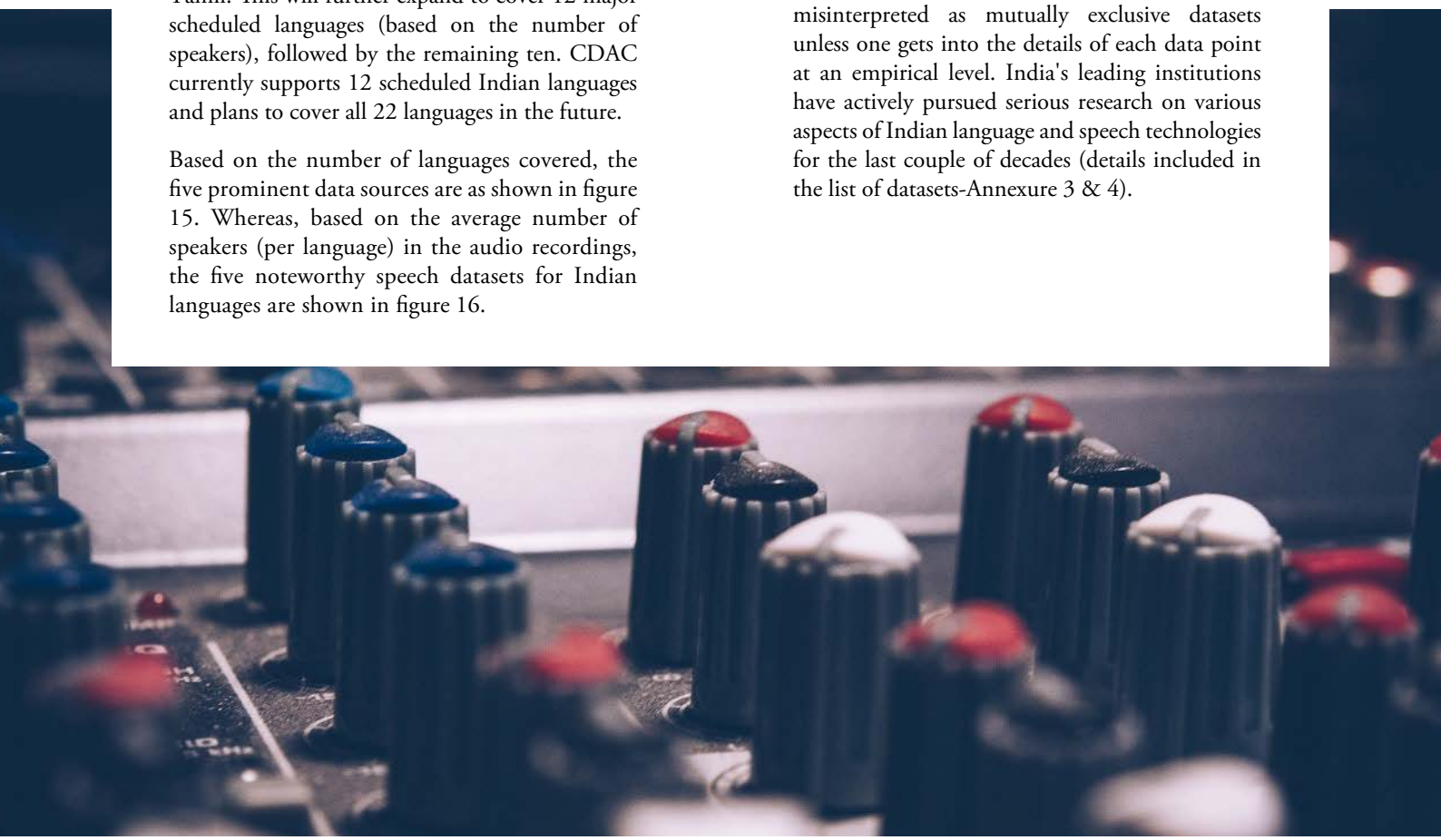
Figure 16: Top 5 datasets based on number of speakers per language (Appendix 3)

NLTM (National Language Translation Mission), also known as 'Bahubashak', is a project that has been initiated by the Prime Minister's Office and is managed by MeitY. This project aims at building technology support systems and products for Indian languages and having them deployed in the field with the help of startups. The pilot project for Bahubashak would involve capturing data in three languages, namely, Hindi, Indian English, and Tamil. This will further expand to cover 12 major scheduled languages (based on the number of speakers), followed by the remaining ten. CDAC currently supports 12 scheduled Indian languages and plans to cover all 22 languages in the future.

Based on the number of languages covered, the five prominent data sources are as shown in figure 15. Whereas, based on the average number of speakers (per language) in the audio recordings, the five noteworthy speech datasets for Indian languages are shown in figure 16.

It needs to be noted that there might be available voice data sources in Indian languages that were not included in this report. In that case, please reach out to the project team of FAIR Forward - Artificial Intelligence for All via fairforward@giz.de.

Also, it is widely observed that there is significant ambiguity in terms of overlapping datasets that may exist in more than one repository and may be misinterpreted as mutually exclusive datasets unless one gets into the details of each data point at an empirical level. India's leading institutions have actively pursued serious research on various aspects of Indian language and speech technologies for the last couple of decades (details included in the list of datasets-Annexure 3 & 4).



"Most of the data used by large companies isn't available to the majority of people. We think that stifles innovation.

So, we've launched Common Voice, a project to help make voice recognition open and accessible to everyone."

Mozilla Open Voice

https://commonvoice.mozilla.org/en

5. Observations

Based on the secondary exploration and the expert opinions of prominent researchers in Indian speech research, the following observations were made on various aspects of Indian-language voice corpora. In general, it was observed that the following challenges might cause obstacles in building a sustainable voice AI ecosystem in terms of developing open speech datasets in Indian regional languages.

5.1 State of Open-Source Voice Data in India

- While LDCIL and TDIL have made significant efforts to bring together the voice and text corpora in Indian languages on a common platform, the adequacy and quality in terms of dialectical coverage and level of transcription are yet to be ascertained. However, in particular for developing general-purpose ASR systems in individual languages, it is not yet adequate.
- It has been observed that the available datasets fall short of the requirements in terms of volume, dialectical coverage, transcription quality, and speaker variations when developing speech recognition systems for specific use cases. Startups need to build their own datasets to ascertain these parameters' robustness rather than rely on the available ones.
- At present, there is hardly any speech recognition research that exploits the characteristics of Indian languages. All the phonemes and models for Indian languages are tuned to existing ones present for English (by adding extra phonemes for Indian languages).
- The existing datasets are not labeled adequately against their properties (like sampling rate, environment, domain, speakers, etc.), making it challenging to offer them under a common platform.

5.2 Working with Indian Languages

- Accents and dialects play a critical role in speech technology. Capturing all possible accents and dialects of a language is a tedious task, especially in India, where almost all the official languages have multiple dialects (e.g., Hindi alone has 48 officially recognised dialects).
- Speech in India is highly colloquial. It is unlikely that an AI speech model might give high accuracy in the real-world scenario as speakers' lack of a disciplined approach might cause systems to fail. Based on speakers' demographics, their native language, and dialect, their sentence constructs may change. Therefore, identifying the right resources or speakers who would provide the data is often a challenging task.
- Out of the 22 languages, little work on voice data collection has been done on two languages - Sanskrit and Sindhi.

5.3 Methods of Speech Data Collection

- A well-articulated voice segment such as a news bulletin on TV or radio may not be of much use for building ASR systems as in everyday speech, people do not pay much attention to articulation and sentence construction as a professional news anchor would do. On top of that, personal factors also play a role, e.g., one's intonation, emotion, etc.
- An empirical challenge with existing datasets collected through various sources and methods is the level and quality of voice transcription. In some cases where there is limited involvement of linguists and domain experts in the planning and transcription of voice datasets, the overall quality of outcome suffers.
- Most of the available corpora are generated through projects targeting domain-specific use cases. Thus, lexicons and observed semantics are also domain restricted. On the contrary, a general-purpose ASR requires a more exhaustive and generic conversational lexicon and corresponding semantics to attain the desired accuracy level.

5.4 Indian Voice Technology Community

- There is no unified information platform even within the organizations. For example, a Delhi branch (or subdivision) would not have enough details about speech data collection going on at any other branch unless it is a joint project.
- If the people working in speech technology (especially data generation) have limited expertise about technical aspects of speech and linguistics, quality data collection, feature extraction, and modeling can become challenging tasks.

Available datasets fall short of the requirements in terms of volume, dialectical coverage, transcription quality, and speaker variations when developing speech recognition systems.

- Startups and small and medium enterprises often consider voice datasets a competitive advantage in the niche market of Indian speech technology. Therefore, it is scarce that they would be willing to make their datasets open.
- There are two schools of thought regarding the approach to achieve the target of generating large enough corpora for efficient ASR and TTS applications in Indian languages:

  Some believe that targeting domain-specific corpora and subsequent aggregation of these voice datasets may help develop a more comprehensive general-purpose corpus for ASR systems. Whereas others believe that creating general-purpose voice datasets should be prioritised from the start. Building a robust general-purpose ASR system is more critical as it can be made domain-specific with minimal extra efforts. The efficacy of expected outcomes from either of the strategies is yet to be ascertained.

5.5 Intent to Open-source

- Certain researchers and organizations dealing with collecting speech data are not independent and are regulated by the project funding agencies' policies. This may sometimes hinder the smooth flow of data and information and make them conservative in sharing their resources with other platforms like ELRA or LDC (Linguistic Data Consortium).
- In consortium projects, public and autonomous organizations work on a single project that involves the collection of speech datasets. When it comes to sharing the dataset in the open domain, each organization has its policies and views, resulting in the data being kept within the consortium and does not become publicly available.
- Government bodies like TDIL and CIIL have often relied on autonomous institutes like IIT Madras, IIIT Hyderabad, or other government bodies like CDAC to collect speech data. An autonomous entity dedicated towards speech data collection may expedited the overall process of data generation.
- It was observed that independent researchers are open to sharing their datasets on any platform to prove useful for the AI community and society at large. All they seek is credit and recognition for their contributions.
- A roadmap is currently being planned by the government (TDIL) to create a common platform where speech corpora collected by different organizations and researchers can be unified and made available as open-source.
- There is no concrete standard agreement for data sharing, which can provide a base for sharing and distributing the datasets across or beyond the participating organizations, with a clear usage policy.

There are two schools of thought when it comes to creating voice datasets for general purpose speech recognition system.

The efficacy of expected outcomes from either of the strategies is yet to be ascertained.

1. One group believes in creating domain specific corpora that can be aggregated into a wider general-purpose corpus for ASR systems.

2. Another group believes in creating general-purpose voice datasets from the start. Building a robust general-purpose ASR system is more important as it can be made domain specific with minimal extra efforts.

Recommendations

1. Platform

A common platform should be built featuring a well-organized mechanism for collecting and maintaining voice datasets. These datasets collected from various sources and organizations must be maintained in a clearly labeled and hierarchical structure that validates the quality of data in terms of their domain, demography of speakers, sources, etc. The searching for datasets based on extended metadata should be a key feature.

The contributors must be encouraged to contribute data towards common data platforms by laying out incentive plans for the contributors wherein they are allowed to use the resources and datasets of the platform for free for a specific period (if not lifetime).

2. Standardization

There is a need for a standard set of guidelines for voice data creation.
This way, all datasets captured from various organizations and researchers meet the same standard and are compatible with each other. Such datasets would be usable for multiple purposes. Multiple efforts have been made to develop a BIS certification type model for datasets, but it has not yet shown any plausible result. Establishing a certifying authority or technical working group for speech datasets might resolve many data compatibility issues in the future.

3. Performance Index

Similar to Google Scholar's H-Index for research publications and its citation, a performance index could be devised for voice datasets to gauge datasets' performance once submitted to a common platform. The said index could be based on the number of times others refer to or use it to train their models. Performance on this index may be considered a criterion for future project allocations to researchers and (or) institutions.

4. Blood Bank Model

A co-creation model where a researcher contributes good quality datasets also gets free access to others' datasets on the platform as an equitable return.

5. Unique Identifiers

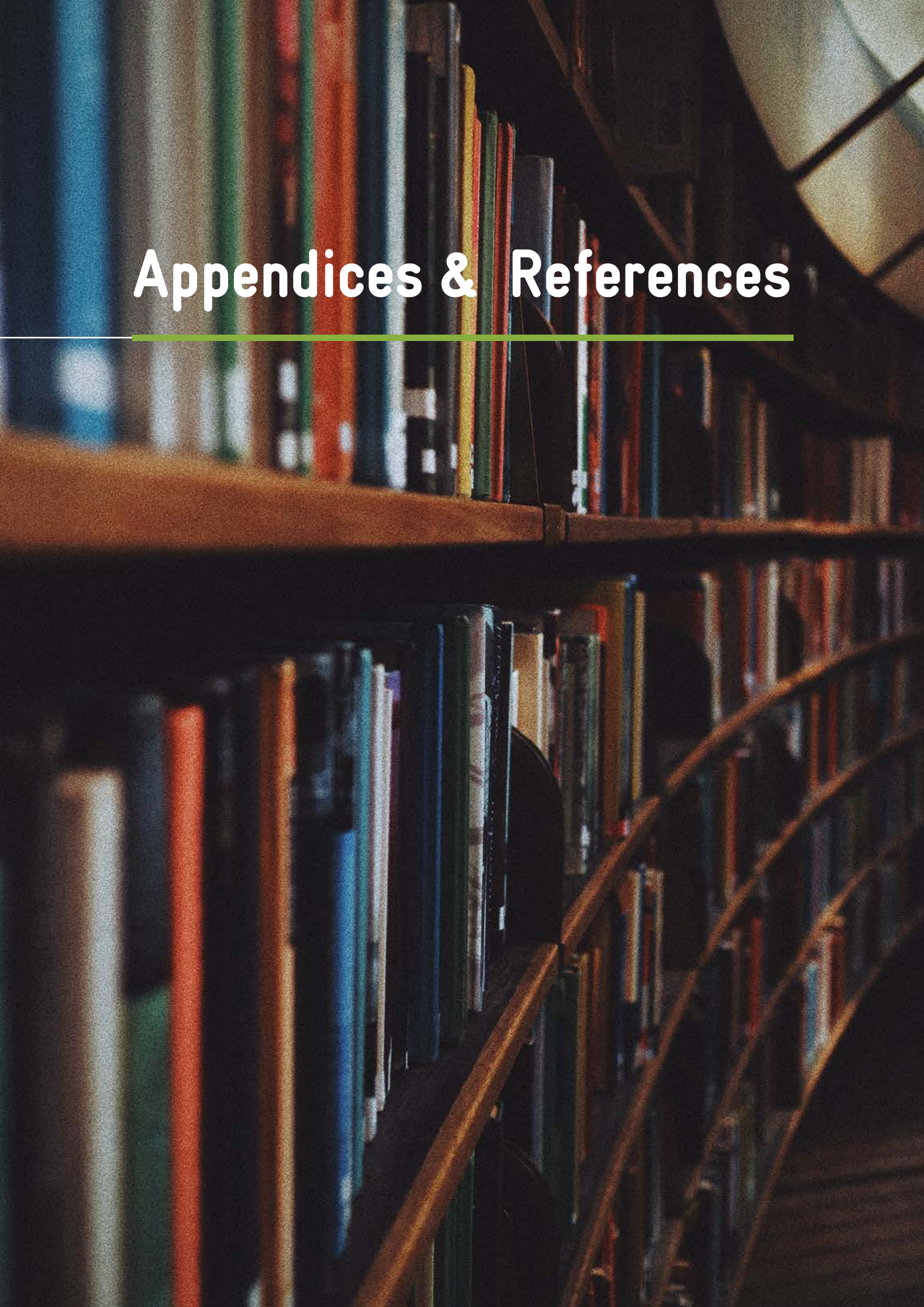
An individual dataset once qualified for inclusion by the common platform should be assigned a unique dataset identification number such as a digital object identifier (DOI). The current issue of ambiguity regarding double-counting the same datasets on multiple platforms could be addressed with this implementation.

6. Other Recommendations

- The building of applications on ASR and (or) TTS, and data collection should go in parallel. Only when we build a speech system will we know the gaps that need to be addressed to gather good quality data. This would lead to formulation of robust standards for data collection.
- Technology should be localised. In India, everybody can benefit from speech systems if these systems focus on users' accents, colloquial ways of speaking, and his/her native language (and dialect), which might not necessarily be one of the official 22 Indian languages. Therefore, efforts must also focus on the collection of voice data for the non-scheduled Indian languages.
- The data that has been collected by using various text sources for read speech generation is required to be crosschecked

against copyright issues in case the data is being commercialised.

- There must be organizational structures and mechanisms to perform manual quality checks (like LDC has) to match the spoken words with the corresponding transcripts. For example, if the speaker pronounces a word differently or adds in some fillers to his speech, quality checks must be put in place to ensure relevant information is added, indicating a particular word being mispronounced or fillers being spoken. Microsoft Research claims to follow manual quality checks for its speech data.
- Efforts must be made to capture good quality general-purpose speech datasets. Such datasets can later be used with some additional changes to adapt to domain-specific requirements, but the vice-versa may not be easy and effective.
- Crowdsourcing should be widely used as it has emerged as an economical and efficient method for generating large volumes of voice data. The volunteers are more natural in their speech when exposed to unsupervised conditions (on a webpage or mobile app) as opposed to a constrained recording environment.



Appendices & References

Appendix 1: List of Contributors

List of experts who participated in qualitative discussions.

Contributors	
Name	**Organization/Institute**
Dr. Aaditeshwar Seth	Gram Vaani
Dr. Anil Kumar Vuppala	IIIT Hyderabad
Prof. A G Ramakrishnan	IISc/Ragavera
Prof. Bharti Ghawali	Dr. Babasaheb Ambedkar Marathwada University, Aurangabad
Dr. Kalika Bali	Microsoft Research India
Mr. Karunesh Arora	CDAC
Mr. Nakul Kundra	Devnagri
Dr. Narayan Choudhary	CIIL/LDCIL
Ms. Nimisha Srivastava	IIIT Hyderabad
Dr. Prasanta Ghosh	IISc
Mr. Roul Nanavati	NavanaTech
Prof. Suryakanth V Gangashetty	KL University
Mr. Udbhav Tiwari	Public Policy Advisor, Mozilla Corporation
Dr. Vivek Raghavan	Ekstep Foundation
Prof. Yegnanarayana	IIIT Hyderabad

Appendix 2: Abbreviations

List of abbreviations used in the report

Abbreviations	
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
TTS	Text to Speech
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit Gmbh
BMZ	German Federal Ministry for Economic Cooperation and Development
IPA	International Phonetic Alphabet
OpenSLR	Open Speech and Language Resources
ELRA	European Language Resources Association
LDC	Linguistic Data Consortium
LDCIL	Linguistic Data Consortium for Indian Languages
CIIL	Central Institute for Indian Languages
CMU	Carnegie Mellon University
AIR	All India Radio
IIT	Indian Institute of Technology
IIIT	International Institute of Information Technology
IVR	Interactive Voice Response
PM	Prime Minister
A2D	Analog to Digital
SNR	Signal to Noise Ratio
TDIL	Technology Development for Indian Languages
ROC	Receiver Operating Characteristic

Appendix 3: List of Datasets

List of speech datasets in Indian languages and their quantitative details.

Datasets					
S. No.	**Database**	**Languages covered**	**Vol Hours (HH.MM)**	**Number of Speakers**	**Speech System**
1.	LDCIL Data Portal	Bengali	128.46	476	ASR
		Bodo	176.53	456	
		Hindi	118.4	489	
		Kannada	179.32	656	
		Konkani	156.37	504	
		Maithili	72.02	300	
		Malayalam	164.01	458	
		Manipuri	156.28	620	
		Marathi	89.17	307	
		Nepali	87.14	350	
		Punjabi	101.09	467	
		Telugu	22.43	80	
		Urdu	99.18	499	
2.	TDIL Speech Corpus	Telugu	-	1073	ASR
		Tamil	-	1000	
		Marathi	-	1500	
		Bengali	-	1000	
		Assamese	-	1023	
		Punjabi	-	-	
3.	NPLT (National Platform for Language Technology)	Assamese	43	2 (1 Male, 1 Female)	ASR/TTS
		Bengali	36+	1500 + 2 (1 Male, 1 Female)	
		Indian English	-	1500	
		Bodo	20	1 (Female)	TTS
		Gujarati	40	2 (1 Male, 1 Female)	
		Hindi	40+	1500+2 (1 Male, 1 Female)	
		Kannada	13+	2 (1 Male, 1 Female)	
		Malayalam	35	2 (1 Male, 1 Female)	
		Manipuri	40+	2 (1 Male, 1 Female)	
		Marathi	31+	2 (1 Male, 1 Female)	
		Oriya	27+	2 (1 Male, 1 Female)	
		Rajasthani	29+	2 (1 Male, 1 Female)	
		Tamil	43	2 (1 Male, 1 Female)	
		Telugu	33	2 (1 Male, 1 Female)	
4.	SCEHMA speech database (Independent Research, BAMU)	Hindi	1.05 (sentences)	30 (18 Male, 12 Female)	ASR
		Marathi	17.36 isolated. 20.29 sentence	500 (300 Male, 200 Female)	
		Indian English	1.23 (sentences), 0.57 (isolated words)	15 (12 Male, 3 Female)	

S. No.	Database	Languages covered	Vol Hours (HH.MM)	Number of Speakers	Speech System
5.	Indian Languages Corpus for Speech Recognition (RP) (CDAC Noida, CDAC Kolkata and KIIT Gurugram)	Hindi Indian English Bengali	210	1500 1500 1524	ASR
6.	Telugu Naturalistic Emotional Speech Corpus (IIIT-Hyderabad)	Telugu	15	38	ASR
7.	Indic: Text to Speech Corpus (RP) (IIIT Hyderabad)	Hindi Malayalam Bengali	25.6 29.1 22.3	1 (for each language)	TTS
8.	IIITH-ILSC Speech Database for Indian Language Identification (IIIT Hyderabad)	Assamese Bodo Dogri Gujarati Hindi Kannada Kashmiri Konkani Maithili Malayalam Manipuri Marathi Nepali Oriya Punjabi Sanskrit Santali Sindhi Tamil Telugu Urdu Indian English	4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5 4.5	50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50	ASR
9.	Common Voice Dataset (Mozilla)	Tamil Assamese Punjabi Oriya	20 0.4 0.13 3	- - - -	ASR

S. No.	Database	Languages covered	Vol Hours (HH.MM)	Number of Speakers	Speech System
10.	Academic Research (Consortium leader – IIIT Hyderabad)	Assamese	12.4	–	–
		Bengali	9.91	–	
		Gujarati	9.71	–	
		Hindi	10.96	–	
		Kannada	10.08	–	
		Malayalam	10.08	–	
		Manipuri	5.31	–	
		Marathi	7.84	–	
		Oriya	9.81	–	
		Punjabi	15.43	–	
		Telugu	10.43	–	
		Urdu	10.8	–	
11.	Microsoft Speech Corpus	Telugu	50	–	ASR
		Tamil	50	–	
		Gujarati	50	–	
12.	All India Radio (News Division)	Multiple Languages	> 1000	1 (per news bulletin)	ASR/TTS
13.	Mann Ki Baat (PM – Narendra Modi)	English	Approx. 30	2 (PM, 1 caller)	ASR/TTS
		Hindi	Approx. 30	2 (PM, 1 caller)	
		Kannada	Approx. 30	2 (PM, 1 caller)	
		Telugu	Approx. 30	2 (PM, 1 caller)	
		Marathi	Approx. 30	2 (PM, 1 caller)	
		Manipuri	Approx. 30	2 (PM, 1 caller)	
		Urdu	Approx. 30	2 (PM, 1 caller)	
		Bengali	Approx. 30	2 (PM, 1 caller)	
		Malayalam	Approx. 30	2 (PM, 1 caller)	
		Tamil	Approx. 30	2 (PM, 1 caller)	
		Assamese	Approx. 30	2 (PM, 1 caller)	
		Oriya	Approx. 30	2 (PM, 1 caller)	
14.	CMU INDIC database (Language Technologies Institute, Carnegie Mellon University)	Hindi	–	1	TTS
		Bengali	–	1	
		Gujarati	–	3	
		Kannada	–	1	
		Marathi	–	2	
		Punjabi	–	1	
		Tamil	–	1	
		Telugu	–	3	

S. No.	Database	Languages covered	Vol Hours (HH.MM)	Number of Speakers	Speech System
15.	CMU Wilderness Multilingual Speech Dataset (Language Technologies Institute, Carnegie Mellon University)	Hindi Fiji Hindi Caribbean Hindustani Awadhi Bengali Kannada Chhattisgarhi Falam Kurukh Magahi Maithili Marathi Malayalam Oriya Tamil Telugu Urdu Sherpa	17.49 19.11 22.18 17.49 22.05 22.48 14.14 22.19 13.17 16.51 13.4 21.52 2.44 7.54 21.14 14.46 19.09 7.12	Generally, 1 speaker per chapter	ASR/TTS
16.	Faith Comes by Hearing Dataset	62 Regional Languages	–	Generally, 1 speaker per chapter	ASR/TTS
17.	Gram Vaani Dataset (ELRA – European Language Resources Association)	Hindi	130	4000 (Bihar, Jharkhand, &Madhya Pradesh (20–25 percent female, 60 percent ‹ 30 years, rural))	ASR
18.	OpenSLR Dataset	Malayalam Tamil Telugu Kannada Gujarati Marathi	5.3 7.04 5.42 8.28 7.53 3.01	42 50 47 59 36 9	TTS
19.	LibriVox Audiobooks Dataset	Bengali Hindi Marathi Oriya Tamil Telugu Urdu	0.03+ 2.1 0.17 0.16 12.2 0.55 1.16	Generally, 1 speaker per audiobook	ASR/TTS

S. No.	Database	Languages covered	Vol Hours (HH.MM)	Number of Speakers	Speech System
20.	SMC Dataset (Swathanthra Malayalam Computing)	Malayalam	1.38	75	TTS
21.	KIIT Speech Dataset (sponsored by Nokia Research Centre, China)	Hindi	› 85	100	ASR/TTS
22.	Multi-variability (MV) Speaker Recognition Database (IIT-Guwahati)	Multiple Indian Languages	-	200	ASR
23.	GlobalPhone Corpus (ELRA + Karlsruhe Institute of Technology (KIT))	Tamil	-	49	ASR
24.	Hindi ASR Challenge Dataset (NLTM + IIT Madras)	Hindi	50	-	ASR
25.	ASR – Hindi (VIT, IISc)	Hindi	-	30	ASR
26.	LDC- Language Data Consortium	Malto	8	27	-
27.	LILA Hindi Belt database	Hindi	-	2023	ASR
28.	LILA Hindi-L1 database (native speakers) LILA Hindi-L1 (non-native)	Hindi	-	3800	ASR
29.	Hindi-English Code-Switching Corpus (IIT-Guwahati)	Hindi-English	25+	101	ASR

Appendix 4: Dataset Details & Source Links

Institutions / Organizations / Programs Active in Voice Research

Datasets			
Sr.	**Database**	**Description**	**Sources/URL**
1.	LDC-IL (Linguistic Data Consortium for Indian Languages)	LDC-IL (Linguistic Data Consortium for Indian Languages)	https://www.ldcil.org/resourcesSpeechCorp.aspx
2.	LDCIL Data Portal	Domain: Contemporary Text (News), Creative Text, Sentence, Date Format, Command and Control Words, Person Name, Place Name, Most Frequent Word - Part, Most Frequent Word - Full Set, Phonetically Balanced, Form and Function – Word Government Organization/Program	https://data.ldcil.org/speech/speech-raw-corpus&att_id=12 https://data.ldcil.org/upload/pubs/RawSpeechOverview.pdf
3.	TDIL Speech Corpus	Domain: Agriculture Government Organization/Program	http://tdil-dc.in/index.php?option=com_download&task=fsearch&lang=en
4.	NPLT (National Platform for Language Technology)	Environment (for non-TTS databases): Studio, Office-Home and Roadside Government Organization/Program	https://nplt.in/demo/index.php?route=product/category&path=75_61
5.	SCEHMA speech database (Independent Research, BAMU)	Domain: Agricultural, polyclinic, accent identification and general speech recognition domain Academic Institute	https://www.springerprofessional.de/en/scehma-speech-corpus-of-english-hindi-marathi-and-arabic-languag/17558080
6.	Indian Languages Corpus for Speech Recognition (RP) (CDAC Noida)	Environment: Office/Home, Roadside, Studio Domain: Most common words, phonetically rich, Agriculture, Travel Age group: 15-30, 31-55, >55 yrs. Government Organization	https://ieeexplore.ieee.org/document/9041171
7.	Telugu Naturalistic Emotional Speech Corpus (IIIT-Hyderabad)	Domain: Films, Short films, Soap Operas Academic Institute	http://speech.iiit.ac.in/svldownloads/IIIT-HTEMD/iiithtemd.html
8.	Indic: Text to Speech Corpus (RP) (IIIT Hyderabad)	Domain: Newspaper Readings Gender: Female (Hindi, Malayalam), Male (Bengali) Academic Institute	https://www.aclweb.org/anthology/2020.lrec-1.789.pdf

Sr.	Database	Description	Sources/URL
9.	IIITH-ILSC Speech Database for Indian Language Identification (IIIT Hyderabad)	Domain: Prasaar Bharati, All Indian Radio, TED-talks, conversational, recorded speeches speech from broadcasts Academic Institute	https://www.isca-speech.org/archive/SLTU_2018/abstracts/Ravi1.html
10.	Common Voice Dataset (Mozilla)	Read Speeches (crowdsourcing)	https://commonvoice.mozilla.org/en/datasets
11.	Academic Research (Consortium leader - IIIT Hyderabad)	Domain: General Purpose Consortium Project	-
12.	Microsoft Speech Corpus	Domain: Multiple	https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e
13.	All India Radio (News Division)	Domain: News	http://newsonair.com/RNU-NSD-Audio-Archive-Search.aspx
14.	Mann Ki Baat (PM - Narendra Modi)	Domain: Social Issues, Economy, Development, India, PMOI Speech	https://www.narendramodi.in/mann-ki-baat#0
15.	CMU INDIC database (Language Technologies Institute, Carnegie Mellon University)	Academic Institute	http://festvox.org/h2r_indic/ http://festvox.org/cmu_indic/index.html
16.	CMU Wilderness Multilingual Speech Dataset (Language Technologies Institute, Carnegie Mellon University)	Domain: New Testaments from Bible.is Academic Institute	http://www.festvox.org/cmu_wilderness/
17.	Faith Comes by Hearing Dataset	Domain: Bible NPO	https://www.faithcomesbyhearing.com/audio-bible-resources/recordings-database
18.	Gram Vaani Dataset (ELRA - European Language Resources Association)	Domain: Local policies, Local news, Agriculture, Health and Social norms, Poetry	http://catalog.elra.info/en-us/repository/browse/ELRA-S0405/
19.	OpenSLR Dataset	Domain: - Wikipedia - Organic handcrafted sentences - Template based - Real-world sentences from TTL applications	http://www.openslr.org

Sr.	Database	Description	Sources/URL
20.	LibriVox Audiobooks Dataset	Domain: Poetry, Human Rights, English Phonetic Pronunciation, Spiritual, Social Drama, Ghazals	https://librivox.org/search?primary_key=0&search_category=language&search_page=1&search_form=get_results
21.	SMC Dataset (Swathanthra Malayalam Computing)	Domain: Multiple	https://releases.smc.org.in/msc-reviewed-speech/ https://github.com/smc/msc https://blog.smc.org.in/malayalam-speech-corpus/
22.	KIIT Speech Dataset (sponsored by Nokia Research Centre, China)	Domain: Messages used in mobile communication	http://www.lrec-conf.org/proceedings/lrec2012/pdf/1132_Paper.pdf
23.	Multi-variability (MV) Speaker Recognition Database (IIT-Guwahati)	Domain: Multiple	https://iitg.ac.in/eee/emstlab/SRdatabase/introduction.php https://www.iitg.ac.in/eee/emstlab/SRdatabase/SRpublications/1.pdf
24.	GlobalPhone Corpus (ELRA + Karlsruhe Institute of Technology (KIT))	Domain: Thinaboomi Tamil Daily newspaper	http://catalog.elra.info/en-us/repository/browse/ELRA-S0205/ http://www.cs.cmu.edu/~tanja/Papers/schultz_icslp02.pdf
25.	Hindi ASR Challenge Dataset (NLTM + IIT Madras)	Domain: Newspaper readings	https://sites.google.com/view/asr-challenge https://github.com/Speech-Lab-IITM/Hindi-ASR-Challenge#link-to-the-data http://tdil-dc.in/index.php?option=com_content&view=article&id=169&lang=en
26.	ASR – Hindi (VIT, IISc)	Domain: Multiple	http://reports.ias.ac.in/report/20583/automatic-speech-recognition---hindi
27.	LDC- Language Data Consortium	Domain: Multiple	https://catalog.ldc.upenn.edu/LDC2012S04
28.	LILA Hindi Belt database	Domain: Mobile telephone conversation	http://metashare.elda.org/repository/browse/lila-hindi-belt-database/3c315548de6e11e2b1e400259011f6ea66b53fb99477493e9deda22e6af0062d/
29.	LILA Hindi-L1 database (native speakers) LILA Hindi-L1 database (non-native speakers)	Domain: Multipl	http://metashare.dfki.de/repository/browse/lila-hindi-l1-database/1d4e20a8de7711e2b1e400259011f6ea1cd92f2a3e7d41559608dff234b78dc1/ http://www.lrec-conf.org/proceedings/lrec2008/pdf/278_paper.pdf
30.	Hindi-English Code-Switching Corpus (IIT-Guwahati)	Domain: Multiple	https://www.iitg.ac.in/eee/emstlab/HingCoS_Database/HingCoS.html

References

1. Office of The Registrar General, India (2011). Census of India 2011. Retrieved from https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf. Pg. 4

2. Times of India. India has just 2.4 percent of world's land but 18 percent of its population. https://timesofindia.indiatimes.com/city/nagpur/india-has-just-2-4-of-worlds-land-but-18-of-its-population/articleshow/69848388.cms

3. Department of Higher Education. Ministry of Education. Government of India. Language education. Retrieved from https://www.education.gov.in/en/language-education

4. Department of Higher Education. Ministry of Education. Government of India. Language. Retrieved from https://www.mhrd.gov.in/hi/sites/upload_files/mhrd/files/upload_document/languagebr.pdf

5. College of Liberal Arts and Sciences. Department of Linguistics. About Hindi. Retrieved from https://linguistics.illinois.edu/hindi/about-hindi

6. Slator. KPMG report examines India's online language and content preferences. Retrieved from https://slator.com/demand-drivers/kpmg-report-examines-indias-online-language-and-content-preferences/

7. Datareportal. Digital 2020: India. Retrieved from https://datareportal.com/reports/digital-2020-india. Pg. 8-24

8. Analytics India Magazine and Jigsaw Academy (2020). State of Artificial Intelligence in India – 2020. Retrieved from https://analyticsindiamag.com/report-state-of-artificial-intelligence-in-india-2020

9. Business Wire. Artificial Intelligence for Speech Recognition Market in India 2019. Retrieved from https://www.businesswire.com/news/home/20200221005349/en/Artificial-Intelligence-for-Speech-Recognition-Market-in-India-Expected-to-Grow-with-a-CAGR-of-65.17-Over-the-Forecast-Period-2019-2024---ResearchAndMarkets.com

10. Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. "LDC-IL Raw Speech Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 160-174.

11. Arora, Sunita & Arora, Karunesh & Roy, Mukund & Agrawal, Shyam & Murthy, B.K.. (2016). Collaborative Speech Data Acquisition for Under Resourced Languages through Crowdsourcing. Procedia Computer Science. 81. 37-44. 10.1016/j.procs.2016.04.027.

12. Panda, Soumya & Nayak, Ajit & Rai, Satyananda. (2020). A survey on speech synthesis techniques in Indian languages. Multimedia Systems. 26. 10.1007/s00530-020-00659-4. Pg. 2

13. Kumar, Munish & Singh, Amitoj. (2019). ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. Artificial Intelligence Review. 10.1007/s10462-019-09775-8. Pg. 1

14. Towards Data Science. Speech recognition is hard. Retrieved from https://towardsdatascience.com/speech-recognition-is-hard-part-1-258e813b6eb7

15. Medium. Deep learning for speech recognition. Retrieved from https://medium.com/@ODSC/deep-learning-for-speech-recognition-cbbebab15f0d

16. Understood. Text-to-speech technology: what it is and how it works. Retrieved from https://www.understood.org/en/school-learning/assistive-technology/assistive-technologies-basics/text-to-speech-technology-what-it-is-and-how-it-works

17. Explain That Stuff. Speech synthesizers. Retrieved from https://www.explainthatstuff.com/how-speech-synthesis-works.html

18. Transcribe Me. Speech recognition training: using annotated data to improve machine learning. Retrieved from https://www.transcribeme.com/speech-recognition-training-annotated-data-improves-machine-learning

19. Sodimana, Keshan & Silva, Pasindu & Sarin, Supheakmungkol & Kjartansson, Oddur & Jansche, Martin & Pipatsrisawat, Knot & Ha, Linne. (2018). A Step-by-Step Process for Building TTS Voices Using Open-source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. 66-70. 10.21437/SLTU.2018-14.

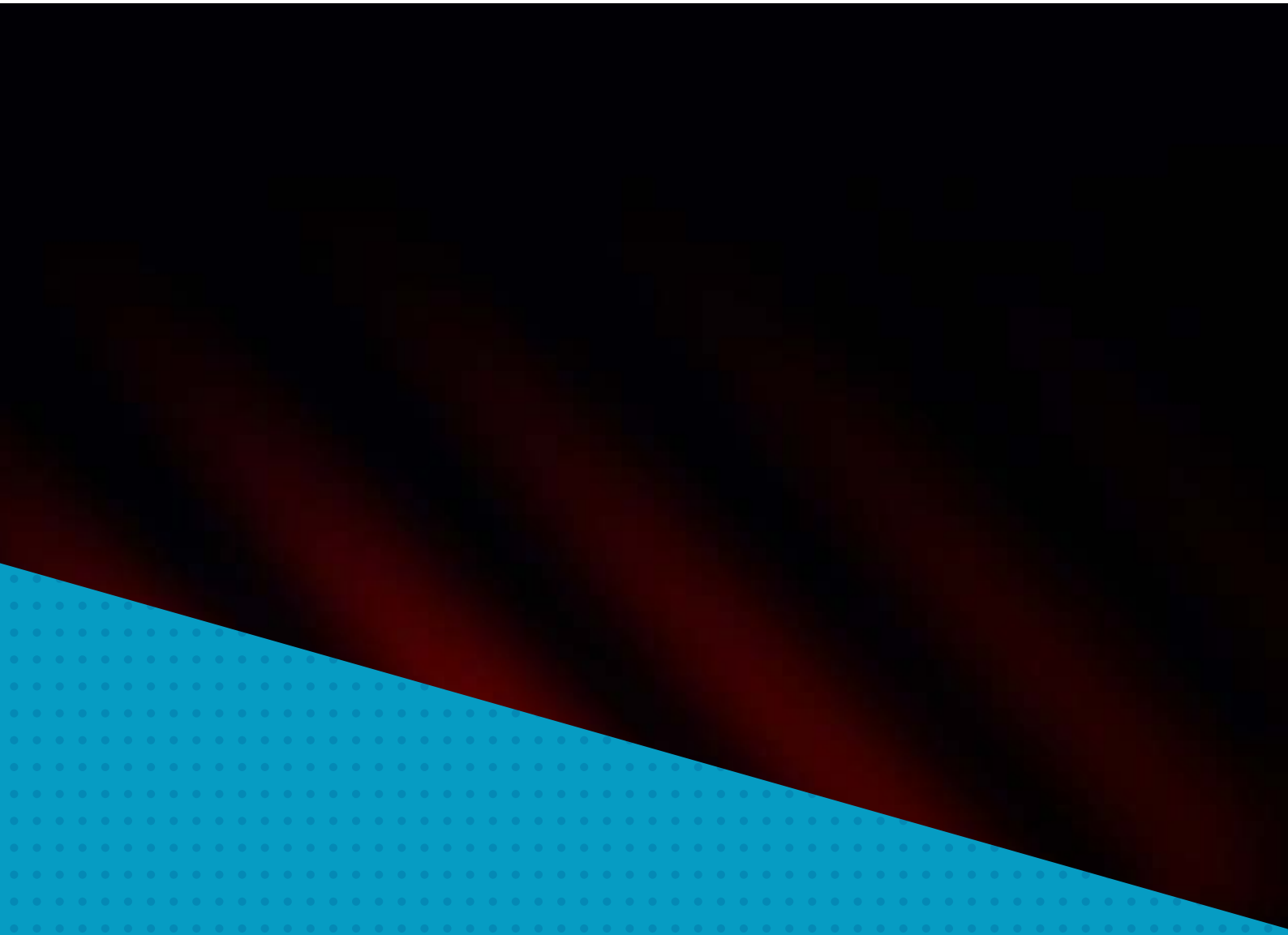
20. Global Market Insights. Smart speaker market size by intelligent virtual assistant. Retrieved from https://www.gminsights.com/industry-analysis/smart-speaker-market

21. Allied Market Research. Smart speaker market size by intelligent virtual assistant. Retrieved from https://www.alliedmarketresearch.com/smart-speaker-market

22. Cudoo. Which countries have the most English speakers. Retrieved from https://cudoo.com/blog/which-countries-have-the-most-english-speakers/

23. Markets and Markets. Smart speaker market with COVID-19 impact analysis via IVA. Retrieved from https://www.marketsandmarkets.com/Market-Reports/smart-speaker-market-44984088.html#:~:text=The percent20smart percent20speaker percent20market percent20is,Units) percent20between percent202020 percent20and percent202025

24. Globalme. Language Support in Voice Assistants Compared. Retrieved from https://www.globalme.net/blog/language-support-voice-assistants-compared/



Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

A2/18 Safdarjung Enclave
New Delhi-110029 India

T: +91-11-49495353
E: fairforward@giz.de
www.giz.de/India