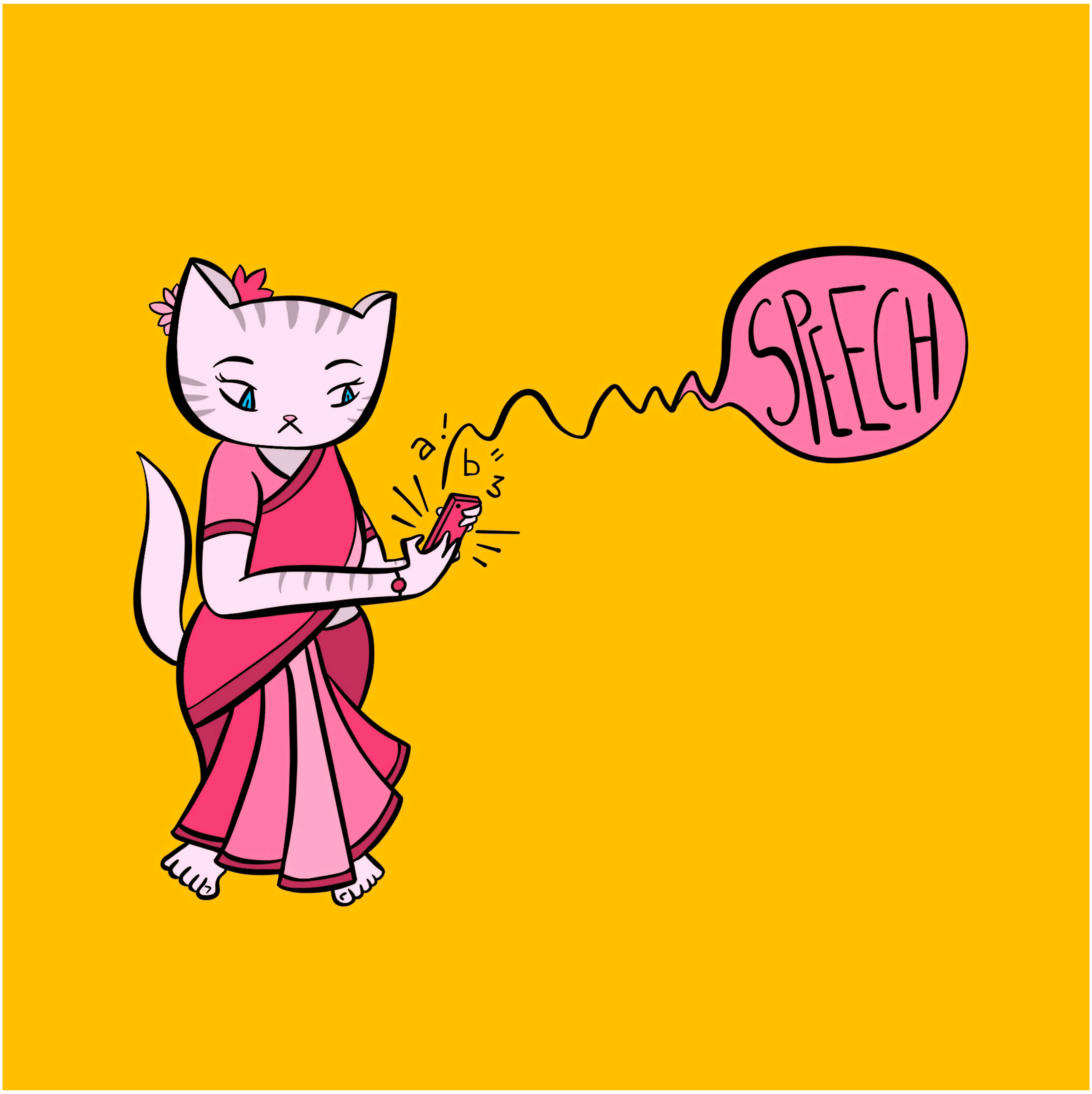

MAKING VOICES HEARD

INDIC TTS

Case Study

Making Voices Heard
Case Study: Indic TTS

Research and Writing **SHWETA MOHANDAS, SAUMYAA NAIDU**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**
Copyediting **THE CLEAN COPY**
Illustration **KRUTHIKA N.S.**
Report Layout and Design **SAUMYAA NAIDU**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. About

“The amount of work in the speech domain for Indian languages is comparatively lower than that for other languages.”[1]

The Indic TTS consortium was created and funded by the Department of Electronics and Information Technology, Ministry of Communications and Information Technology,[2] Government of India, to create more Indic language speech data to reduce the data divide between English and Indian languages. The Indic TTS website describes this as “a project on developing text-to-speech (TTS) synthesis systems for Indian languages, improving quality of synthesis, as well as small footprint TTS integrated with disability aids and various other applications”.[3] In a recently published paper on voice technologies, researchers involved in the TTS project stated that the paucity of content in Indian languages was stark when it came to the multimedia domain and digital assistants,[4] thereby highlighting the need for speech data in Indian languages. This paucity was attributed to the lack of localisation of technologies like optical character recognition (Optical character recognition or OCR is the electronic or mechanical conversion of images of typed, handwritten, or printed text into machine-encoded text in a way that can be read by speech-to-text systems)[5],neural machine translation(neural machine translation uses computing systems that mimic the working of the human brain to predict the order of words in sentences)[6] and text-to-speech systems.

The speech data for the database was collected through the joint efforts of the 13 consortium members: IIT Madras, IIIT Hyderabad, IIT Kharagpur, IISc Bangalore, CDAC Mumbai, CDAC Thiruvananthapuram, IIT Guwahati, CDAC Kolkata, CDAC Pune, SSNCE Chennai, DA-IICT Gujarat, IIT Mandi, and PESIT Bangalore. The database and text-to-speech synthesisers were built for 13 languages, namely, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odia, Rajasthani, Tamil, and Telugu.

2. Methodology and process

2.1. Language and text selection

The process of creating voice datasets in Indian languages involved several steps, beginning with the selection of languages that are the focus of the project and

[1] Baby, Arun et al., “Resources for Indian Languages”, In *Proceedings of CBBLR workshop, International Conference on Text, Speech and Dialogue*. Springer, 2016.

[2] “Indic TTS”, *Indic TTS*, https://www.iitm.ac.in/donlab/tts/; Department of Electronics and Information Technology (DEITY) has been renamed to Ministry of Electronics and Information Technology (MEITY) 03 November 2021.

[3] “Indic TTS”, *Indic TTS*. accessed 3 November 2021.

[4] Baby Arun, “Resources for Indian Languages”.

[5] “An Introduction to Optical Character Recognition for Beginners”, *Towards Data Science*, accessed 5 January 2022, https://towardsdatascience.com/an-introduction-to-optical-character-recognition-for-beginners-14268c99d60

[6] Tan,Zhixing et al., “Neural machine translation: A review of methods, resources, and tools”, *AI Open Volume 1*.(2020):5-21, https://doi.org/10.1016/j.aiopen.2020.11.001.

then building speech technologies using the voice datasets. The selection of the 13 languages was based on the following criteria: optimal text selection, speaker selection, pronunciation variation, recording specification, text correction for handling out-of-the-vocabulary words, and data verification.[7] To ensure the quality of data, characteristics that affect speech synthesis quality such as encoding (converting one form of data to another), sampling rate (number of samples of audio recorded every second) etc. were considered. The sentences for the speech recordings were taken through web crawlers from newspaper reports, Wikipedia pages, websites, and blogs in the respective Indian language. To achieve good coverage of topics and words, sentences were also taken from different types of literature, including children's stories, science writing, tourism content, etc. Care was also taken to ensure that the texts were commonly used, free of errors, easy to read, and covered a wide range of words and syllables. Code-mixed sentences were avoided.

2.2. Speaker selection and recording

To create speech recordings for the datasets, two voice talents – a male and a female – were chosen for each language. The recordings were made in a studio room without noise or echo for clarity of the recordings. The voice talents were voice professionals who were either voice artists or newsreaders to ensure clarity in the pronunciation and diction. They were given breaks every 45 minutes to avoid fatigue. In each recording, individual sentences were isolated. A total of 40 hours of speech data was collected for a given language – 20 hours of Indian monolingual/single language data (10 hours each of male and female voice data) and 20 hours of English data recorded by first language speakers (10 hours each of male and female voice data). The recorded files were stored in .wav format to ensure that the recordings were of high sound quality.

2.3. Text-to-speech synthesis

One of the researchers in the Indic TTS project defines text-to-speech synthesis as the "process of converting an arbitrary input text to its corresponding speech output". In the context of Indian languages, the TTS system uses syllables or phonemes (units of sound that can distinguish one word from another in a particular language) as a sub-word unit (where words are split into smaller words that occur more frequently). The three major components involved in building a TTS system are text parsing, speech segmentation, and speech modelling.[8] Simply put, the objective of a TTS system is to convert text into speech output. TTS systems can be divided into two types – domain-specific and vocabulary independent. In the case of domain specific systems, the words/text to be synthesised should be limited to a particular domain, such as banking or railway broadcast, while for vocabulary independent systems, any text will work.

[7] Indic TTS", *Indic TTS*.

[8] Baby, Arun, "A Unified Approach to Speech Synthesis in Indian Languages", (MS Thesis, IIT Madras, 2019), 1–93, https://www.arunbaby.com/assets/docs/MSthesis_2019.pdf.

3. Languages

The main motivation for the project was to address the unavailability of voice data in Indian languages. The functioning of a TTS system is dependent on the training data that is fed into the system, which includes speech .wav files along with a transcript of the corresponding text. The TTS project aims to develop text-to-speech synthesisers for 13 Indian languages, which could help researchers and developers work on Indian voice applications. One of the goals of the project is to make the voice of the text-to-speech system sound as natural and understandable as possible. The first phase of the project concentrated on three languages (3 Indo-Aryan languages and 3 Dravidian languages), the second phase added 7 more languages to the study.

4. Access and accessibility

The TTS project was started with the idea of giving people with disabilities access to regional information on the internet, such as news reports in Indian languages. Since the consortium is a publicly funded project, the datasets and research have been made public on its website. The datasets are available free of cost to researchers – they just need to log in to the website to use them. Start-ups and businesses that want to use the data can sign a Memorandum Of Understanding with Indic TTS and access the data.

5. Privacy and data collection

As stated earlier, the text data for training the systems was taken from publicly available sources such as online news portals, Wikipedia pages, websites, and blogs; hence, privacy and data protection are not significant concerns. Additionally, with regard to the speech data, the readings were done by professional voice artists who recorded sounds and words for the project based on a script provided to them by the researchers.

6. Challenges

One of the main challenges for the researchers was ensuring that the datasets were comprehensive and accurate while keeping the cost of creating and accessing them low. Since the project was publicly funded, the researchers needed to work with the available funding and ensure that the research was accessible and free. As stated earlier, the data and the research are open to researchers, and start-ups can request the data after signing an MOU. Another challenge was making the speech output sound more human-like and less robotic, similar to the heavily funded and data-rich interfaces of Amazon and Google. The other challenge was making the output speech systems context-specific, such as with children’s books.

7. Future of Indic TTS

The project looks at continuing research and data collection with the help of government funding. Given the scale and amount of funding needed for such projects, including the requirement of infrastructure and trained human resources, the government is the primary source of funding. With the new funding from the Ministry of Electronics and Information Technology, the researchers at IITM have started a project to make English lecture videos available in Indian languages. The objective of this project is to make lectures in different domains, like humanities, healthcare, etc., freely accessible to students in their languages. This is a small-scale project, and Indic TTS hopes to expand it to more languages and subjects.

Disclaimer: This is an independent case study conducted as a part of the Making Voices Heard Project, supported by the Mozilla Corporation. The researchers have not received any external remuneration as a part of this case study, and claim no conflict of interest.