Automatic Indic Speech Transcription

ZEN3 INFOSOLUTIONS
A Tech Mahindra Company

Traditional, human-centric approaches to speech transcription require huge allotments of skilled professional hours. Consequently, automated speech transcription technology has the potential to save huge amounts of time and energy while simultaneously expanding the array of practicable speech-to-text applications.

Automated transcription doesn't just save cost relative to human translators-

It enables the collection of vast quantities of speech data in a text format that's far easier to store, analyze, and integrate with broader datasets.

Until recently, the downside of automatic transcription was dramatically reduced accuracy. Skillful human transcription involves understanding how to correctly write huge vocabularies with widely varying pronunciations, dialect, and bi-lingual usage patterns. Automatic transcription tools need deep rules for understanding these subtleties. And this rules-based understanding needs to be repeatedly developed for each human spoken language.

Today, cutting-edge automated approaches are beginning to rival the accuracy of human transcribers for the first time= with dramatically improved speed and cost. Intensive efforts are required, however, to bring these new techniques to bear on all the globes languages.

Our Client

Our client, a widely recognized innovator and one of the top names in tech industry, is a leading pioneer in automated speech transcription technology. Their speech engine achieves transcription accuracy almost at par with human transcription.
While this feat was achieved on an English language corpus, the client sought to expand its speech capabilities to Indic languages. They selected Zen3 to oversee the annotation of Indic texts for this effort, a core part of the project's data foundation.

Objectives

Our team was tasked with the collection and annotation of conversational data in various Indic languages and dialects over a demanding time frame, preparing the speech data for comparison to other languages and ultimate inclusion in speech-to-text applications.

Annotation refers to the detailed labeling of text transcripts at a granular level involving specific topics and grammatical functions. This information is used for improved functionality in successive iterations of the automatic speech recognition tool. Please see the brief Appendix at the end of this case study for an example of lingual annotation.

Our Services

The project involved collection, transcription and annotation of ~5000 hours of conversational speech in Indic languages including Telugu, Kannada, and Tamil, among others. This ambitious effort required generating a random mix of conversa- tional topics and then setting uniform guidelines for annotation. The planning for the task of annotating and contextualizing the collected data included creating pronunciation dictionaries, entity tagging, emotion labeling, and developing lexicons, among other policy systems.

Speech data was transcribed in a semi-automated fashion. A low resource speech model was built using a small set of manually transcribed speech (for every language). This model was used to bootstrap the transcription effort for the remain- ing voice samples. The transcribers' work was streamlined to focus on correcting mistakes committed by the automated speech recognition tools. Efficiency and accuracy metrics also improved due to Zen3's speech technology, which spotted transcription errors on the fly for immediate rectification.

Project Challenges

The huge number of Indic languages and dialects is rendered even more complex by extensive inter-lingual (L1/L2) influence and bi-lingual speech patterns. This fact of India's rich history and localized culture renders the task of collecting and annotating this speech corpus more demanding than English, widely spoken in only a few relatively centralized versions worldwide. Special attention was required when developing annotation rules for these multi-directional inter-lingual influences.

The demanding timeframe of this project necessitated a proactive management approach founded on proven best practices, detailed below.

Best Practices and Solutions

Zen3 has the right mix of speech, language and technical capabilities necessary for tackling this challenge. Our experience with semi-automated transcription and annotation pipelines ensured the fulfillment of quality and speed objectives.

Automation tools, custom built by Zen3, streamlined the most repetitive parts of the process to drive far greater efficiency for tasks where human review is essential.

First, multiple data sources were used to collect conversational speech samples through an in-house voice recorder which recorded over VOIP/PSTN networks. 8-element array microphones were used to gather in-person conversation. These conversations were further diarized using beam-forming techniques to separate sources. Audio files were then automatically segmented to be annotated by transcribers in the next phase.

Annotation quality-control measures were deployed to automatically identify predictable transcription errors. Any errors identified were sent to a manual quality control check.
Transcription is aided by a low-resource speech recognition model that semi-automates the process while retaining human supervision. The process results in a new acoustic/language model built from the labeled/validated data.

Key Outcomes

Crediting the efficiency and output quality of Zen3's work, client expanded Zen3's role to perform similar work for North Indian languages like Hindi, Gujarati among others.

- **Pronunciation dictionaries for Indic dialect specific utterances**
- **Speech transcription efficiency improved by 32%**
- **Speech model with % WER was one of the deliverables boosting the confidence in the transcribed data**
- **Accuracy levels of 96% achieved after only 2 QC passes**
- **Region/L2 specific lexicon rules for borrowed words**

Languages	Native Language Speech Collected (Hrs)	English Speech by Native Language Speakers Collected (Hrs) (Specific to Region)	Regions/Dialects	Cities Collected the Speech from
Hindi	400	200	North Zone North East Zone East Zone Central Zone West Zone	Chandigarh Lucknow Patna Bhopal, Raipur Jaipur
Tamil	600	400	Chennai Tamil North Tamil Central Tamil Madurai Tamil Kongu Tamil Nellai Tamil	Chennai Krishnagiri and Dharmapuri Thanjavir and Dindigul Madurai and Coimbatore Triunelveli, Kovailpatti
Telugu	400	200	Telangana Telugu Rayalaseema Telugu Coastal Andhra Telugu	Chennai Krishnagiri and Dharmapuri Thanjavir and Dindigul Madurai and Coimbatore Triunelveli, Kovailpatti
Marathi	400	200	Aagri Standard Varhadi East	Mumbai Pune, Nashik and Kolhapur Nagpur
Bengali	250	100	Varhali Standard (Rarhi) Jharkhandi	Akola, Amravati Kolkata Bardhaman, Medinipur
Kannada	350	180	Badagu Northern Coastal Dialects South Karnataka	Ooty, Nilgiri Dharwad Mangalore Bangalore
Malayalam	390	220	Malabar Travancore Kasargode	Kannur, Wayanad, Palakkad Trissur, Kollam Trivandrum, Kasargode, Thallasery

Entity Tagging

English

Hello. Huh hi, [prs::hum-] Kishore [-prs::hum] . <SIL> How are you?<SIL> Yes. Yeah
previously <FIL/> [loc::city-] <LM>Bangalore</LM> [-loc::city] having
lot of <FIL/> greenery
[loc::reg-] Forests [-loc::reg] are <FIL/> we are cutting down

Telugu

[prs::hum-] [-prs::hum]
[loc::reg-] [-loc::reg]

Sentiment Tagging

English

[sentiment=--] wow what a fantastic topic [-sentiment=-surprised, happy]
due to the increase in global warming [sentiment=--] it's actually affecting all the
people's health [-sentiment=-empathy]
[sentiment=--] see no matter how much we deny to accept it [-sentiment=-angry] like

Telugu

[sentiment=--] ?
[-sentiment=-surprised]
[sentiment=--]
[-sentiment=-unhappy]

Lexicon Rules across Indic Languages

Indian languages have rules which do not allow for certain phone sequences to occur. These phone sequences would gener- ally occur in the case of borrowed words like for example from English

G2P Rules Example

At the graphemic level, Bengali has three sibilant graphemes {শ ষ স} which ideally should correspond to /ʃ/, /ʂ/ and /s/ respectively. At the phonemic level however, we do not find them contrasting. At the phonemic level there is only one phoneme /ʃ/ representing the sibilant sounds in the language. /s/ occurs as an allophone of /ʃ/ in the following contexts:

1. Word initial position in combination with [p], [pʰ], [t], [tʰ], [k], [kʰ], [m], [n], [ɾ], [l].
Eg: [spɔɾʃo] ‘touch’; [spʰoʈik] ‘crystal’; [stɔb] ‘hymn’; [stʰɑn] ‘place’; [skɔndɦo] ‘shoulder’; [skʰɔlon] ‘falling’; [smito] ‘sweet smile’; [sneɦo] ‘affection’; [sɾiʃʈi] ‘creation’; [slok] ‘a couplet’.

2. Word medial position in combination with [t], [tʰ], [n], [ɾ], [l].
Eg: [dɔstɑnɑ] ‘gloves’; [ɑstʰɑ] ‘faith’; [ɔssnɑto] ‘unbathed’; [osɾu] ‘tear’; [ɔslil] ‘obscene’

ZEN3 INFOSOLUTIONS

A Tech Mahindra Company

Contact Us

Zen3 Infosolutions Private Ltd.
e-mail: info@zen3tech.com