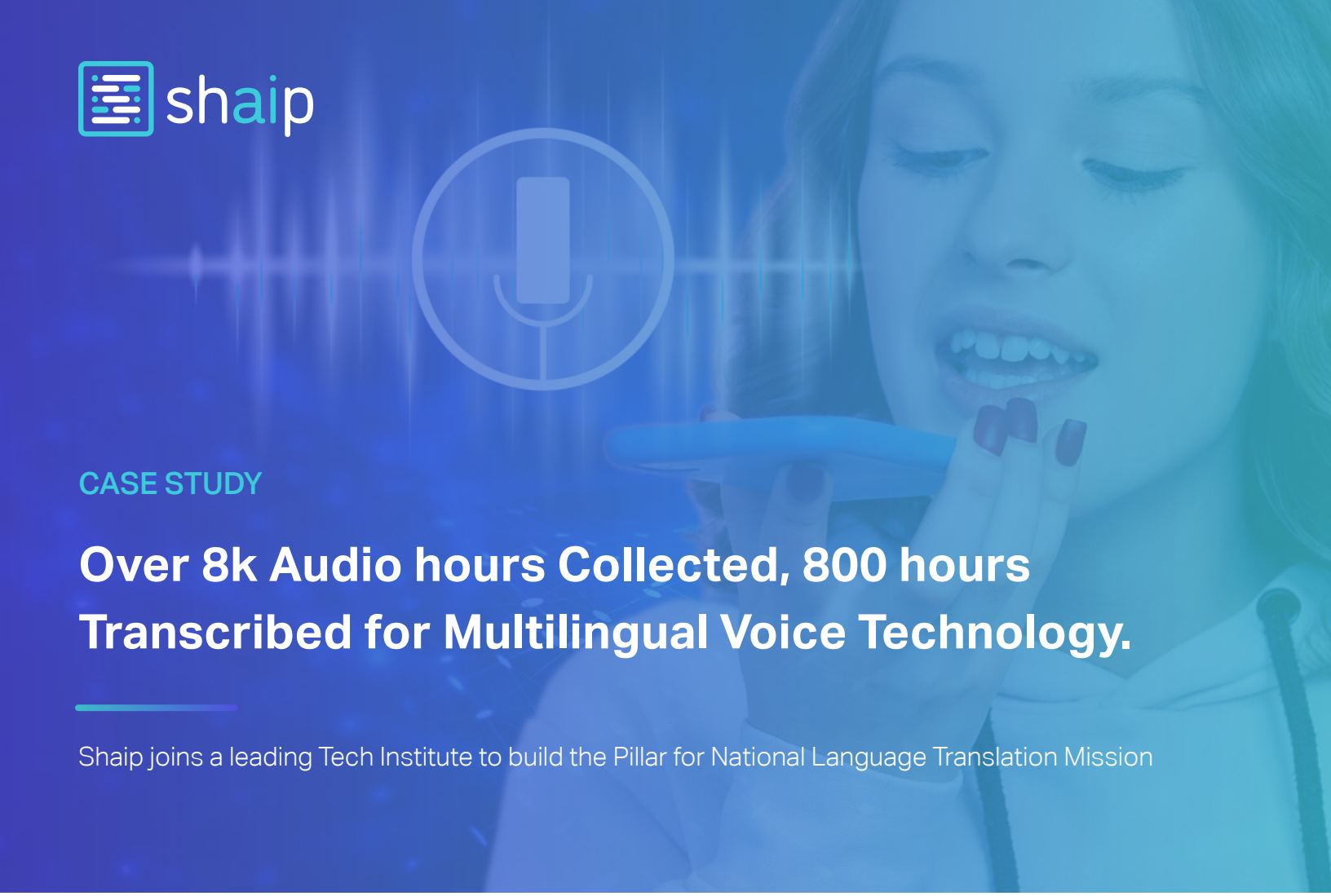
shaip



CASE STUDY

Over 8k Audio hours Collected, 800 hours Transcribed for Multilingual Voice Technology.

Shaip joins a leading Tech Institute to build the Pillar for National Language Translation Mission

COMPANY
A leading Indian Tech Institute

INDUSTRY
Education

USE CASE
Conversational AI: Automatic Speech Recognition

OUR OFFERING
- Audio Data Collection
- Audio Transcription

KEY STATS

Hours of Data Collected
8,000 Hrs

No. of Districts from where audio data was collected
80

Project Timeline
less than 5 months

OVERVIEW

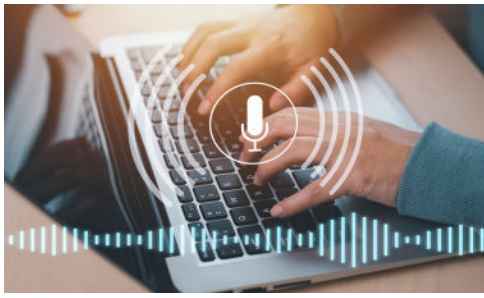
India needed a platform that would concentrate on creating multilingual datasets and AI-based language technology solutions in order to provide digital services in Indian languages. To launch this initiative, The Client partnered with Shaip to collect, and transcribe Indian language datasets to build multi-lingual speech models.

CHALLENGES

To assist the client with their Speech Technology speech roadmap for Indian languages, the team needed to acquire, segment and transcribe large volumes of training data to build AI model.

The critical requirements of the client were:

Data Collection

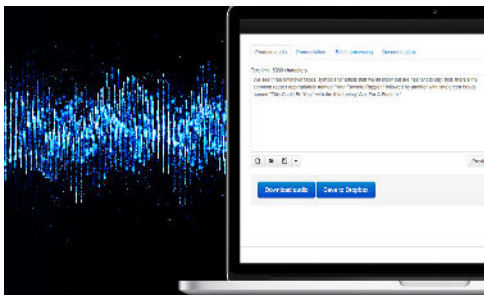


- Acquire 8000 hours of training data from remote locations of India
- The supplier to collect Spontaneous speech from Age Groups of 20-70 years
- Ensure a diverse mix of speakers by age, gender, education and dialects
- Each audio recording shall be at least 16kHz with 16 bits/sample.

Quality Check & Feedback

All recordings to undergo quality assessment and validation, only validated speech recordings to be delivered

Data Transcription



- Follow details transcription guidelines around Characters and Special Symbols, Spelling and Grammar, Capitalization, Abbreviations, Contractions, Individual Spoken Letters, Numbers, Punctuations, Acronyms and Initialisms, Disfluent Speech, Unintelligible Speech, Non-Target Languages, Non-Speech

SOLUTION

With our deep understanding of conversational AI, we helped the client collect, transcribe the audio data with a team of expert collectors, linguists and annotators to build large corpus of audio data from remote parts of India.

The scope of work for Shaip included but was not limited to acquiring large volumes of audio training data, transcribing the data and delivering corresponding JSON files containing the metadata [for both speakers and transcribers. For each speaker, the metadata includes an anonymized Speaker ID, device details, demographic information like gender, age, and education, along with their pincode, socio-economic status, languages spoken, and a record of their life's stay

duration. For every transcriber, the data incorporates an anonymized Transcriber ID, demographic details similar to the speakers', their transcription experience duration, and a thorough breakdown of languages they can read, write, and speak.

Shaip collected 8000 hours of audio data / Spontaneous speech at scale and transcribing 800 hours, while maintaining desired levels of quality required to train speech technology for complex projects. Explicit Consent Form was taken from each of the participants. The / Spontaneous speech collected were on based on University-provided images. Of 3500 images, 1000 are generic and 2500 relate to district-specific culture, festivals, etc. Images depict various domains like train stations, markets, weather, and more.

1. Data Collection

State	Districts	Audio Hrs	Transcription Hrs
Bihar	Saran, East Champaran, Gopalganj, Sitamarhi, Samastipur, Darbhanga, Madhepura, Bhagalpur, Gaya, Kishanganj, Vaishali, Lakhisarai, Saharsa, Supaul, Araria, Begusarai, Jahanabad, Purnia, Muzaffarpur, Jamui	2000	200
Uttarpradesh	Deoria, Varanasi, Gorakhpur, Ghazipur, Muzzaffarnagar, Etah, Hamirpur, Jyotiba Phule Nagar, Budaun, Jalaun	1000	100
Rajasthan	Nagaur, Churu	200	20
Uttarakhand	Tehri Garhwal, Uttarkashi	200	20
Chhattisgarh	Bilaspur, Raigarh, Kabirdham, Sarguja, Korba, Jashpur, Rajnandgaon, Balrampur, Bastar, Sukma	1000	100
West Bengal	Paschim Medinipur, Malda, Jalpaiguri, Purulia, Kolkatta, Jhargram, North 24 Parganas, Dakshin Dinajpur	800	80
Jharkhand	Sahebganj, Jamtara	200	20
AP	Guntur, Chittoor, Visakhapatnam, Krishna, Anantapur, Srikakulam	600	60
Telangana	Karimnagar, Nalgonda	200	20
Goa	North+South Goa	100	10
Karnataka	Dakshin Kannada, Gulbarga, Dharwad, Bellary, Mysore, Shimoga, Bijapur, Belgaum, Raichur, Chamrajnagar	1000	100
Maharashtra	Sindhudurg, Dhule, Nagpur, Pune, Aurangabad, Chandrpur, Solapur	700	70
Total		**8000**	**800**

General Guidelines

Format:

- Audio at 16 kHz, 16 bits/sample.
- Single channel.
- Raw audio without transcoding.

Style:

- Spontaneous speech.
- Sentences based on University-provided images. Of 3500 images, 1000 are generic and 2500 relate to district-specific culture, festivals, etc. Images depict various domains like train stations, markets, weather, and more.

Recording Background:

- Recorded in a quiet, echo-free environment.
- No smartphone disturbances (vibration or notifications) during recording.
- No distortions like clipping or far-field effects.
- Vibrations from phone unacceptable; external vibrations are tolerable if audio is clear.

Speaker Specification:

- Age range from 20-70 years with balanced gender distribution per district.
- Minimum of 400 native speakers in each district.
- Speakers should use their home language/dialect.
- Consent forms mandatory for all participants.

2. Quality Check & Critical Quality Assurance

The QA process prioritizes quality assurance for audio recordings and transcriptions. Audio standards focus on precise silences, segment duration, single-speaker clarity, and detailed metadata including age and socio-economic status. Transcription criteria emphasize tag accuracy, word veracity, and correct segment details. The acceptance benchmark dictates that if more than 20% of an audio batch fails these standards, it's rejected. For less than 20% discrepancies, replacement recordings with similar profiles are required.

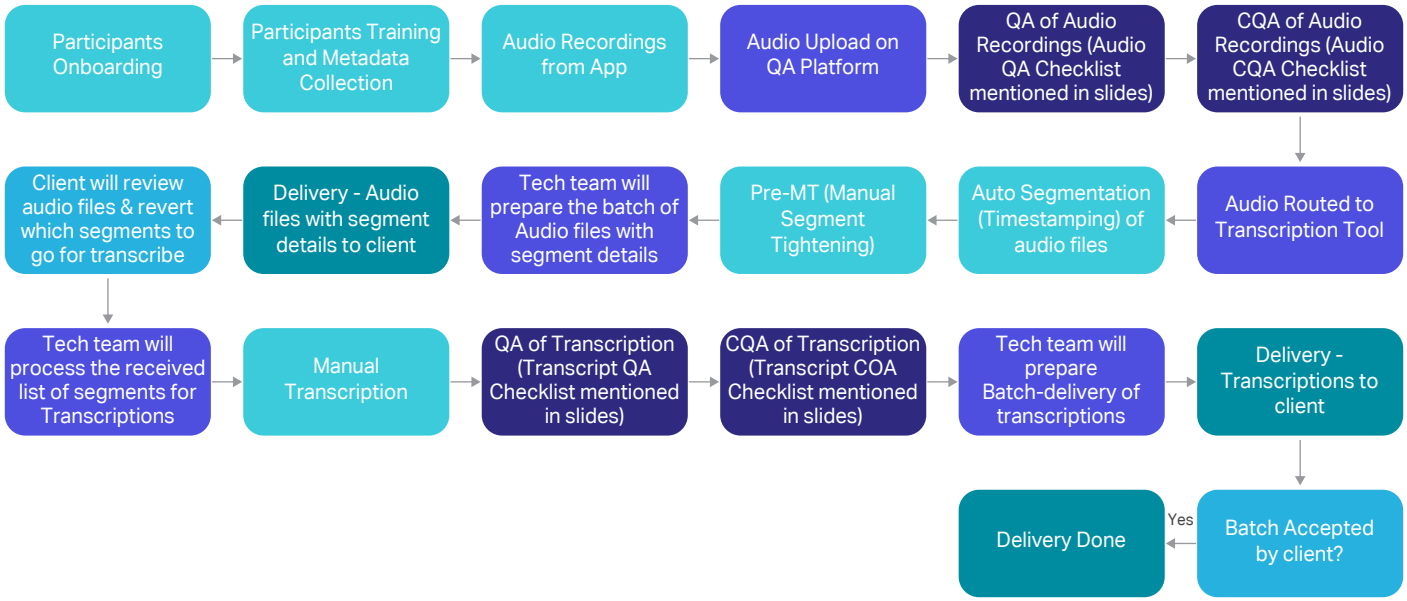
3. Data Transcription

Transcription guidelines emphasize accuracy and verbatim transcription only when words are clear and understandable; unclear words are marked as [unintelligible] or [inaudible] based on the issue. Sentence boundaries in long audio are marked with , and no paraphrasing or correction of grammatical errors is allowed. Verbatim transcription covers errors, slangs, and

repetitions but omits false starts, filler sounds, and stutters. Background and foreground noises are transcribed with descriptive tags, while proper names, titles, and numbers follow specific transcription rules. Speaker labels are used for every sentence, and incomplete sentences are indicated with.

4. Project Workflow

The workflow describes the audio transcription process. It starts with onboarding and training participants. They record audio using an app, which is uploaded to a QA platform. This audio undergoes quality checks and automatic segmentation. The tech team then prepares segments for transcription. After manual transcription, there's a quality assurance step. Transcriptions are delivered to the client, and if accepted, the delivery is deemed complete. If not, revisions are made based on client feedback.



THE OUTCOME

The high-quality audio data from expert linguists will enable our client to accurately train and build multilingual Speech Recognition models in various Indian languages with different dialects in the stipulated time. The Speech recognition models can be used to:

- Overcome language barrier for digital inclusion by connecting the citizens to the initiatives in their own mother tongue.
- Promotes Digital Governance
- Catalyst to form an ecosystem for services and products in Indian languages
- More localized digital content in the domains of public interest, particularly, governance & policy

CUSTOMER TESTIMONIAL

★★★★★

We are in awe of Shaip's expertise in the conversational AI realm. The task of handling 8000 hours of audio data along with 800 hours of transcription across 80 diverse districts was monumental, to say the least. It was Shaip's deep comprehension of the intricate details and nuances of this domain that made the successful execution of such a challenging project possible. Their ability to seamlessly manage and navigate through the complexities of this vast amount of data while ensuring top-notch quality is truly commendable.

shaip

Shaip provides high-quality data across multiple data types (text, audio, image & video) to companies looking to build unbiased and high quality AI/ML models. Shaip licenses, collects and annotates data for Healthcare, Conversational AI, Computer Vision and Generative AI/LLM use cases. Going beyond data, Shaip offers a complete Responsible LLM Toolkit to align, evaluate, and enhance large language models using reinforcement learning from human feedback (RLHF). Headquartered in Kentucky with offices in Silicon Valley and India, our global team blends data science expertise with deep industry knowledge. Visit us at www.shaip.com.