shaip

CASE STUDY

Over 3k hours of Audio Data Collected, Segmented & Transcribed to build Multi-lingual Speech Tech in 8 Indian languages.

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

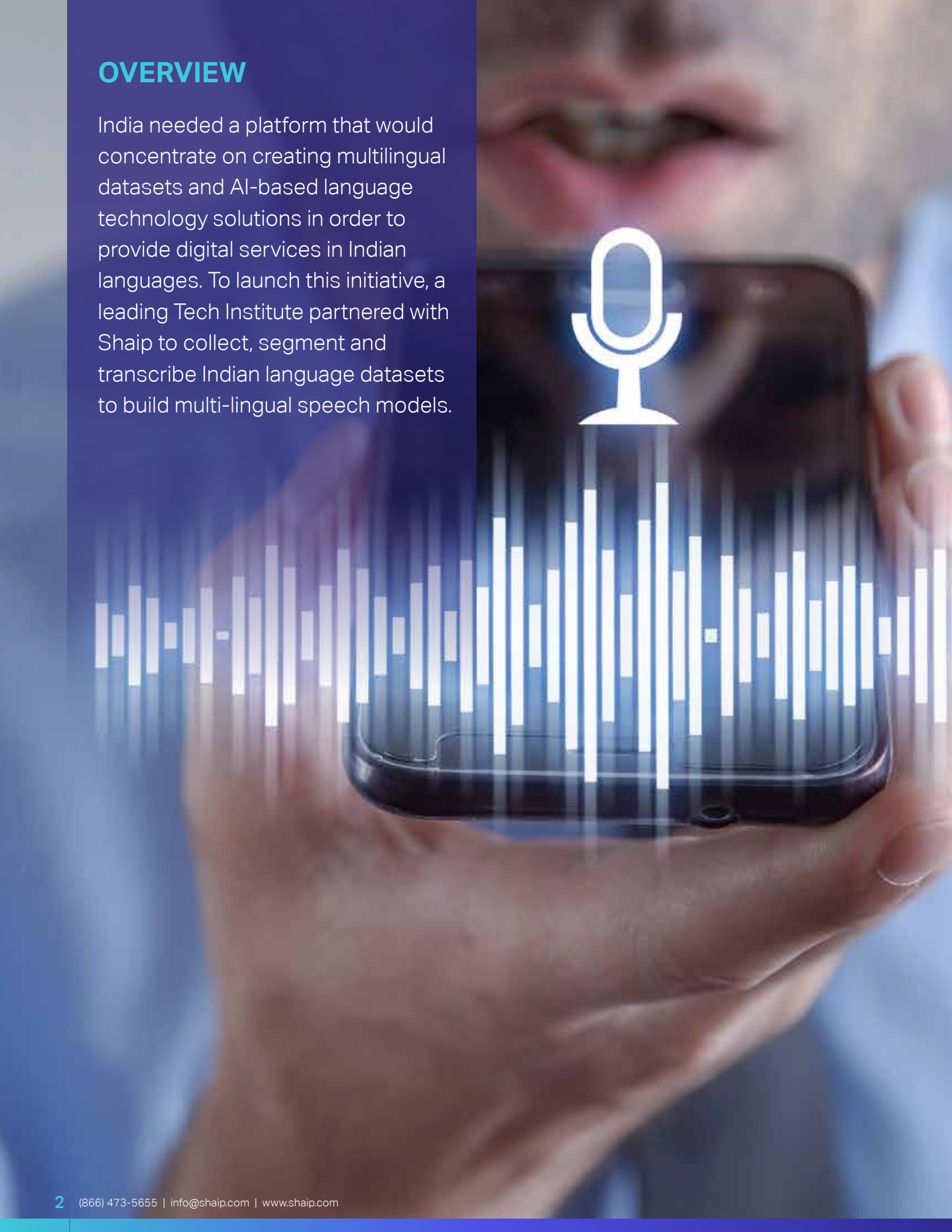
OUR OFFERING
- Audio Data Collection
- Audio Segmentation
- Audio Transcription

KEY STATS

Hours of Data Collected	No. of Languages	Languages Supported	Project Timeline
3,000 Hrs	**8**	**Hindi, Malayalam, Marathi, Gujarati, Punjabi, Assamese, Bengali, and Tamil.**	**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, a leading Tech Institute partnered with Shaip to collect, segment 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 3000 hours of training data in 8 Indian languages with 4 dialects per language.
- For each language, the supplier will collect Extempore Speech and Conversational Speech from Age Groups of 18-60 years
- Ensure a diverse mix of speakers by age, gender, education and dialects
- Ensure a diverse mix of recording environments as per Specifications.
- Each audio recording shall be at least 16kHz but preferably 44kHz

Data Segmentation

- Create speech segments of 15 seconds and timestamp the audio file to the milliseconds for each given speaker, type of sound (speech, babble, music, noise), turns, utterances, and phrases in a conversation
- Create each segment for its targeted sound signal with a 200-400 milliseconds padding at the start and end of a sound signal.
- For all segments, the following objects must be present and filled i.e., Start Time, End Time, Segment ID, Loudness Level, Primary Sound Type, Language code Speaker ID, etc.

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, segment and transcribe the data with a team of expert collectors, linguists and annotators to build large corpus of audio dataset in 8 Indian languages.

The scope of work for Shaip included but was not limited to acquiring large volumes of audio training data, segmenting the audio recordings in multiple, transcribing the data and delivering corresponding JSON files containing the metadata [SpeakerID, Age, Gender, Language, Dialect, Mother Tongue, Qualification, Occupation, Domain, File format, Frequency, Channel, Type of Audio, No. of speakers, No. Of Foreign Languages, Setup used, Narrowband or Wideband audio, etc.]. Shaip collected 3000 hours of audio data at scale while maintaining desired levels of quality required to train speech technology for complex projects. Explicit Consent Form was taken from each of the participants.

1. Data Collection

Sr. No.	Language	No. of Hours	Extempore MonoLingual (80%)		Conversational (2 or more Speakers) (20%)	
			Narrowband (Telephony)	Wideband (Non Telephony)	Narrowband (Telephony)	Wideband (Non Telephony)
1	Hindi	300	90	150	0	60
2	Malayalam	250	75	125	0	50
3	Marathi	500	150	250	0	100
4	Gujarati	500	150	250	0	100
5	Punjabi	500	150	250	0	100
6	Assamese	200	60	100	0	40
7	Bengali	500	150	250	0	100
8	Tamil	250	75	125	0	50
Total		**3000**	**900**	**1500**	**0**	**600**

General Guidelines	Extempore MonoLingual Guidelines	Conversational Guidelines
Diversity • Gender: 50% male, 50% female, +/- 10%. • Minimum 4 dialects per language • Age Groups: 18-60 yrs. **Audio Properties** • Audio format: WAV format ('.wav'). • Sampling Frequency=16Khz • Low Background Noise	A. Collected speech data from 4800 unique speakers in 4 dialects for every language. B. Collected data from domains such as weather, news, entertainment, health, agriculture, education, job, & finance. C. Length did not exceed more than 20 mins.	A. Data collected from multiple speakers on topics such as sports, news, weather, politics, business, govt. schemes etc. based on a draft script. B. Each speaker's conversations lasts around 3-5mins recorded from smartphones C. About 2000-2500 conversations per language collected from 1200-1400 speakers, while maintaining the Male-Female ratio.

2. Data Segmentation

» The audio data that was collected was further bifurcated into speech segments of 15 seconds each and timestamped to the milliseconds for each given speaker, type of sound, turns, utterances, and phrases in a conversation

» Created each segment for its targeted sound signal with a 200-400 milliseconds padding at the start and end of a sound signal.

» For all segments, the following objects were present and filled i.e., Start Time, End Time, Segment ID, Loudness Level (Loud, Normal, Quiet), Primary Sound Type (Speech, Babble, Music, Noise, Overlap), Language Code Speaker ID, Transcription etc.

3. Quality Check and Feedback

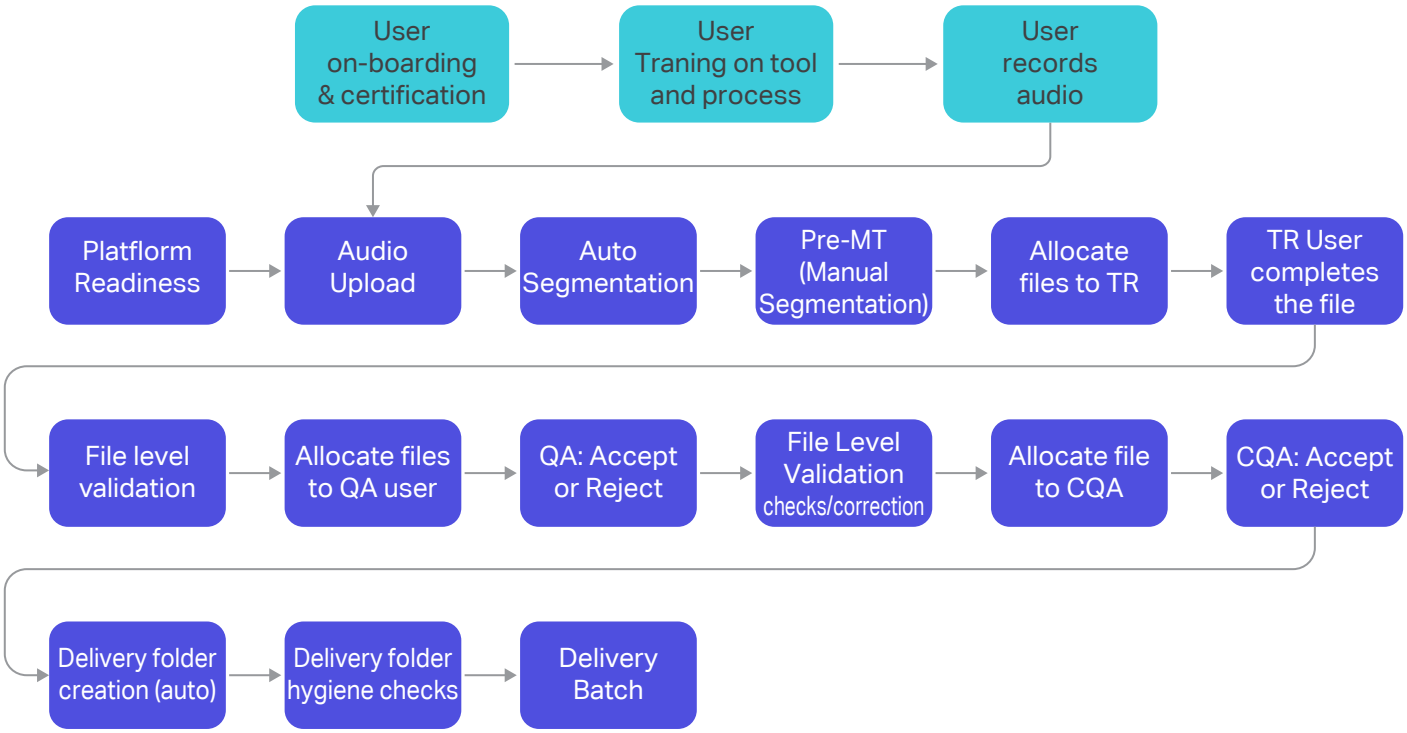
» All recordings were assessed for quality and only validated speech recordings with WER of 90% and TER of 90% were delivered

» Quality Checklist Followed:

- Max 15 seconds of segment length
- Transcription from specific domains, namely: Weather, different types of news, entertainment, health, agriculture, education, jobs or finance
- Low background Noise
- No Audio clip off - No distortion
- Correct audio segmentation for transcription

4. Data Transcription

All spoken words, including hesitations, filler words, false starts, and other verbal tics, were captured accurately in the transcription. We also followed details transcription guidelines around capital and lowercase letters, spelling, capitalization, abbreviations, contractions, numbers, punctuation, Acronyms, Disfluent Speech, non-speech noises etc. Moreover the Work Flow followed for Collection and Transcription is as below:



THE OUTCOME

The high-quality audio data from expert linguists will enable our client to accurately train and build multilingual Speech Recognition models in 8 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 were impressed with Shaip's expertise in conversational AI space. Their overall project execution competency from sourcing, segmenting, transcribing and delivering the required training data from expert linguists in 8 languages within stringent timelines and guidelines; while still maintaining the acceptable standard of quality.

shaip

Headquartered in Louisville, Kentucky, Shaip is a fully managed data platform designed for companies looking to solve their most demanding AI challenge enabling smarter, faster, and better results. Shaip supports all aspects of AI training data from data collection, licensing, labeling, transcribing, & de-identifying by scaling seamless of our people, platform, & processes to develop AI/ML models. To learn more about how to make your data science team and leaders' life more manageable, visit us at www.shaip.com.