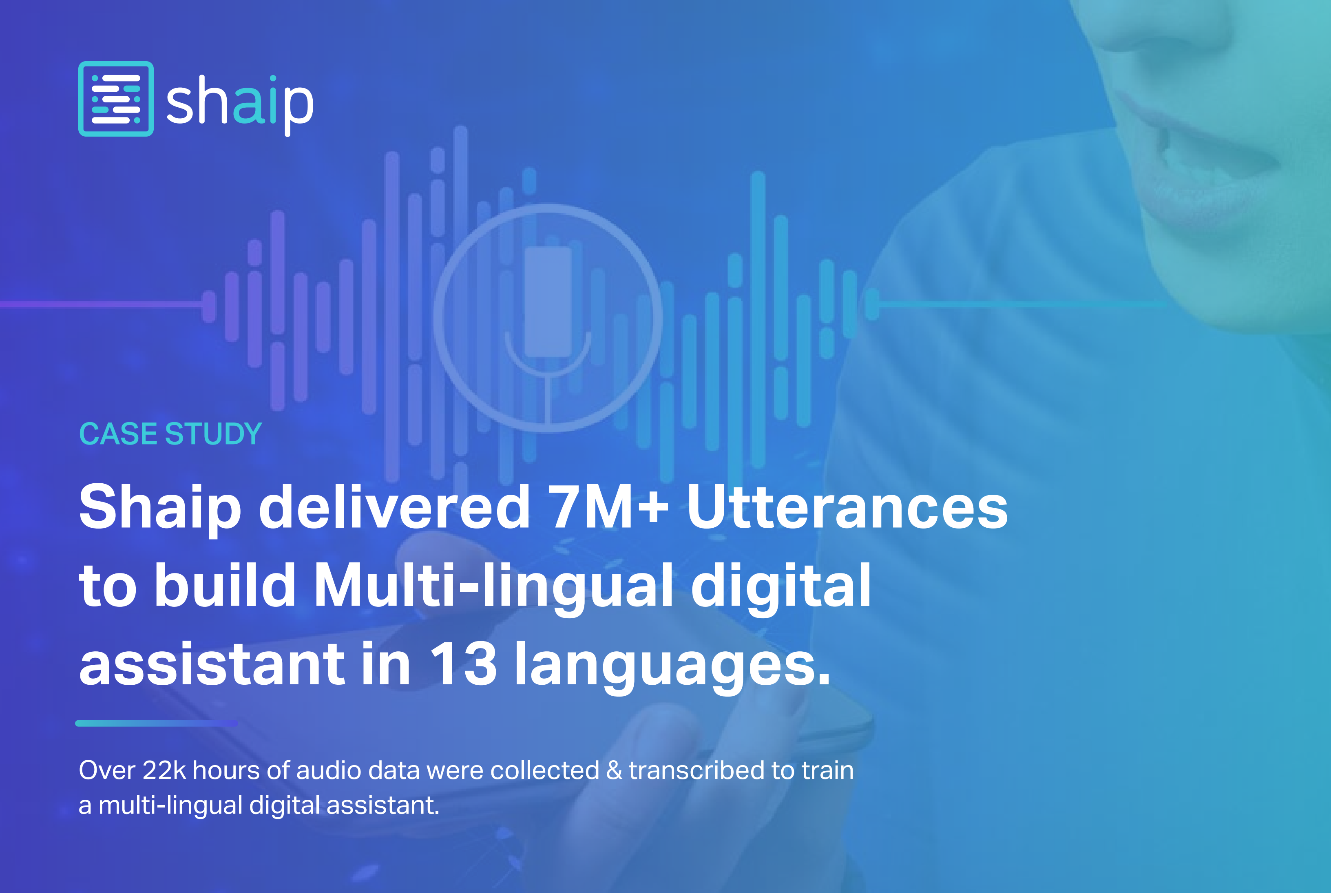


shaip

CASE STUDY

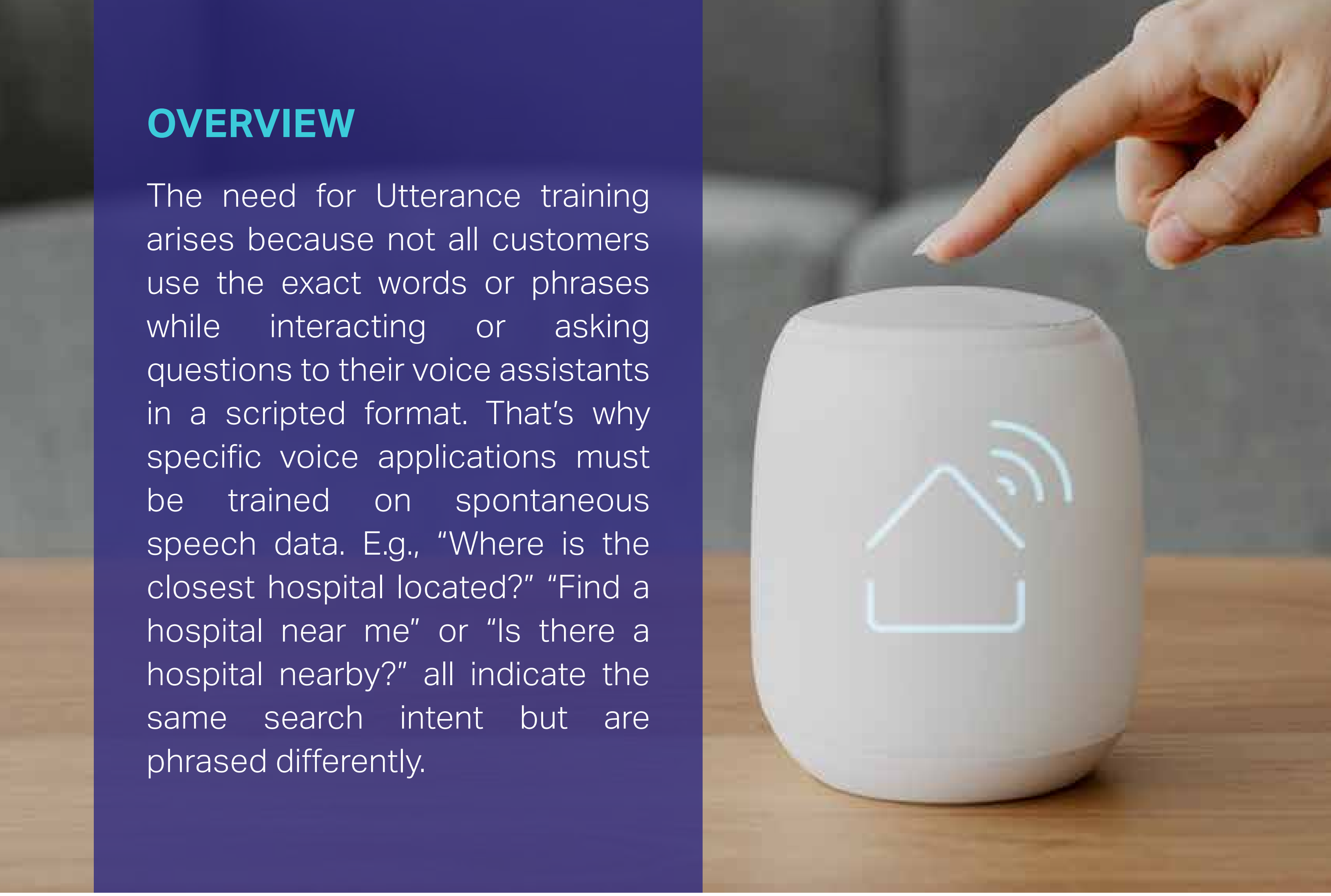
# Shaip delivered 7M+ Utterances to build Multi-lingual digital assistant in 13 languages.

Over 22k hours of audio data were collected & transcribed to train a multi-lingual digital assistant.

| COMPANY | INDUSTRY | USE CASE | OUR OFFERING |
|---|---|---|---|
| A leading Conglomerate | Information Technology | Conversational AI: Utterance Data Collection | • Audio Collection<br>• Audio Transcription<br>• Audio Annotation |

## KEY STATS

| Hours of Data Collected | No. of Languages | No. of Utterances | Languages Supported | Project Timeline |
|---|---|---|---|---|
| **22,250 Hrs** | **13** | **7M+** | **Arabic, Hindi, Chinese, Russian, Japanese, Danish, Korean, Dutch, Spanish, Canadian, Polish, Turkish** | **7-8 months** |



## OVERVIEW

The need for Utterance training arises because not all customers use the exact words or phrases while interacting or asking questions to their voice assistants in a scripted format. That's why specific voice applications must be trained on spontaneous speech data. E.g., "Where is the closest hospital located?" "Find a hospital near me" or "Is there a hospital nearby?" all indicate the same search intent but are phrased differently.

## CHALLENGES

To execute clients' Digital Assistant's speech roadmap for worldwide languages, the team needed to acquire large volumes of training data for the speech recognition AI model.

**The critical requirements of the client were:**

- Acquire large volumes of training data (single speaker utterance prompts of not more than 3-30 seconds long ) for speech recognition services in 13 global languages
- For each language, the supplier will generate text prompts for speakers to record (unless the client supplies) and transcribe the resulting audio.
- Provide audio data and transcription of recorded utterances with corresponding JSON files containing the metadata for all recordings.
- Ensure a diverse mix of speakers by age, gender, education & dialects
- Ensure a diverse mix of recording environments as per Specifications.
- Each audio recording shall be at least 16kHz but preferably 44kHz

# SOLUTION

With our deep understanding of conversational AI, we helped the client collect, transcribe and annotate the data with a team of expert linguists and annotators to train their AI-powered Speech Processing multilingual Voice Suite.

The scope of work for Shaip included but was not limited to acquiring large volumes of audio training data for speech recognition, transcribing audio recordings in multiple languages for all languages on our Tier 1 and Tier 2 language roadmap, and delivering corresponding JSON files containing the metadata. Shaip collected utterances of 3-30 seconds at scale while maintaining desired levels of quality required to train ML models for complex projects.

- Audio Collected, Transcribed & Annotated: **22,250 hours**
- Languages Supported: **13**
- Number of Utterances: **7M+**
- Timeline: **7-8 months**

| Language | No. of Hours | Language | No. of Hours | Language | No. of Hours |
|---|---|---|---|---|---|
| **Danish** | 2000 | **Korean** | 1500 | **Saudi Arabian Arabic** | 1500 |
| **Dutch** | 2000 | **Mainland Chinese** | 2000 | **Taiwan Chinese** | 2000 |
| **French Canadian** | 1000 | **Mexican Spanish** | 1250 | **Turkish** | 1250 |
| **Hindi** | 2000 | **Polish** | 2000 | | |
| **Japanese** | 2000 | **Russian** | 2000 | | |

While collecting audio utterances at 16 kHz, we ensured a healthy mix of speakers by age, gender, education, and dialects in diverse recording environments.

## THE OUTCOME

The high-quality utterance audio data from expert linguists empowered the client to accurately train their multilingual Speech Recognition model in 13 Global Tier 1 & 2 languages in the stipulated time. With gold-standard training datasets, the client can offer intelligent and robust digital assistance to solve future real-world problems.

## CUSTOMER TESTIMONIAL

★★★★★

*“After evaluating many vendors, the client chose Shaip because of their expertise in conversational AI projects. We were impressed with Shaip’s project execution competence, their expertise to source, transcribe and deliver the required utterances from expert linguists in 13 languages within stringent timelines and with the required quality.”*

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