Quick Research on STT/TTS for Indian Languages

1. Objective

Based on the Pranav sir's request to research STT and TTS models for Indian languages, this document explores offline-capable options (i.e., ones you can download and run locally without calling external APIs) and includes small proof-of-concepts to compare trade-offs.

2. Approach

- STT: Ran OpenAI's Whisper (small) model in Google Colab to transcribe a short Hindi audio clip.
- TTS: Tried Indic TTS, but the zip file kept getting corrupted during extraction, so it didn't work. Then attempted to use the Coqui XTTS-v2 multilingual model for Hindi, but ran into PyTorch unpickling issues in this runtime. As a fallback, used espeak-ng to synthesize a sample Hindi sentence.
- Round-trip: Fed the synthesized audio back into Whisper to get a transcription and get a rough sense of intelligibility.

3. Local vs API

- Local/offline: No per-use cost, data stays local, and full control over the pipeline. Setup can be fiddly (e.g., dependency problems or serialization guards like with XTTS-v2).
- API/hosted: Easier to start with, but depends on network, may have costs, and gives less control.
- This version sticks to offline/local proof-of-concepts. An API-hosted baseline (e.g., Whisper API) can be added later to compare speed, quality, and cost.

4. Results

- Input audio: Short Hindi clip- http://hindi.voiceoversamples.com/HIN_M_AbhishekS.mp3

- STT output (Whisper): "इसके बाद हमेक आजे खिलाडी के बाद करेंगे जिसे ज़्ज़िज़्वल्कब में भारत्ये फैंज को काफी मीदे रहेंगी इसके लाडी के बारे में आपको बस इतना बतादो कि यह वो खिलाडी है जिसने अंटरनेशनल क्रिक्ट में अपनी अंट्री का एलान और चेल्यक अगें जोड़दार प्रदशन से किया आज से वल्कब के दारान रोज हम आपको मिलते रहेंगे EAM क्रिक्ट वल्कट तो हैजार सात अप्टेट में

  `Inference time: 111.03 seconds`"

- TTS input: "नमस्ते दुनिया, यह एक टेस्ट है।"

- TTS output: Saved as `output_hindi.wav` (synthesized via `espeak-ng`; note quality is robotic but fully offline).

- Round-trip transcription: Whisper transcribed the synthesized audio as: "अवाश्टे दुल्या या खेख पेस्ठें"

5. Attempted Higher-Quality TTS

Tried Coqui XTTS-v2 for Hindi. Loading was attempted with safe-unpickling tweaks, but PyTorch's serialization safeguards (multiple internal classes needed to be allowlisted) blocked a stable inference in this Colab environment. This remains a follow-up item if a more controlled setup is used.

6. Observations

- Whisper worked out of the box for Hindi and gave a readable transcript, though some parts were incorrect or garbled.
- - XTTS-v2 looked promising but was too fragile to get working and Indic TTS didnt work at all; fallback to espeak-ng ensured the pipeline could complete.
- - espeak-ng is reliable and offline, but the speech sounds robotic, which hurt the round-trip transcription accuracy.
- - The round-trip (TTS→STT) showed that the synthesized audio lost clarity, likely due to the synthetic voice quality.

Google collab link-

https://colab.research.google.com/drive/1fKtWIXSYI1dgZjGyCKfh7KOVCz-V1T-I?usp=sharing