• Voice Recognition uses SST along with speaker recognition

← SFT: (speech to text)

Falls under audio processing

- SST (speech to text) ⇒ input is spoken audio → output: text

  • Goal: convert human speech into text

Steps 1) Audio capture

  • Microphone records audio → sound wave sampled (Ex 16KHz)
  → stored as a wave form

2) Feature Extraction

  • Raw audio is too messy for processing directly so its transformed into Features

  • Common: MFCC (Mel Frequency Cepstral Coefficients) - a way of representing sound that matches human hearing sensitivity

  • think of it as: condensing sound into a compact fingerprint for each small time slice

3) Acoustic Model

  • maps short audio frames → Phonemes (basic sound units in language)

  • old methods: HMMs + GMMs • Modern method: CNN, RNN, transformers

4) Language model

  • helps guess the most probable sequence of words

Ex: acustic model: "recognize speech" raw model output: "wreck a nice to reach"
Then the model picks the more likely option

5) Decoading: combines acoustic and language models to produce Final text

- Moder Advancements

Ex whisper → End-to-End models: skip phonems steps, directly map audio → text using architectures like RNN-transducers or transformers

Multilingual: Model trained on many languages at once

Robustness: Noise suppression, speaker seperation, diarization (who said what)