TTS (Text to sprech)

STT IVR on next page

- TTS (text to sprech) ⇒ input is text → output is spoken audio
  - goal: to convert text into natural (human-like speech

Steps: 1) text analysis / normalizatio
- Convert num, abbreviations etc into sperm form Ex 12.50 → twelve dollar fifty cents

2) Phoneme conversion
- translate words into their phonetic pronunciation

3) Prosody Prediction
- Predict Pitch, Rhythm, and stress patterns to make speech sound natural (instd of robotic)

4) Wave form Generation
- Earlier: Concatenative TTS - stitched together pre recorded human voice samples
- Later: parametric TTS - generated speech from acoustic parameters (less natural sounding)
- Now: Neural TTS - Neural vocoders like WaveNet generate extremly realistic waveforms

- Modern pipeline Example (Tacotron 2 + WaveNet)

1) tacotron 2 - takes text, outputs a mel spectrogram (time vs Frequency plot of audio energy)

2) WaveNET / HiFi GAN - takes that spectrogram, generates realistic audio wave forms