Voice Agent

MACHINE LEARNING ENGINEER IN THE GENERATIVE AI ERA

Lecture Overview

- What is a Voice Agent?
- Introduction to GPT4o-Realtime
- Chained Methods vs End-to-End Models
- How ASR Works & Popular Models (related to Project 3)
- How TTS Works & Popular Models
- Common Audio Data Processing Pipelines (e.g., Emilia)

What is a Voice Agent?

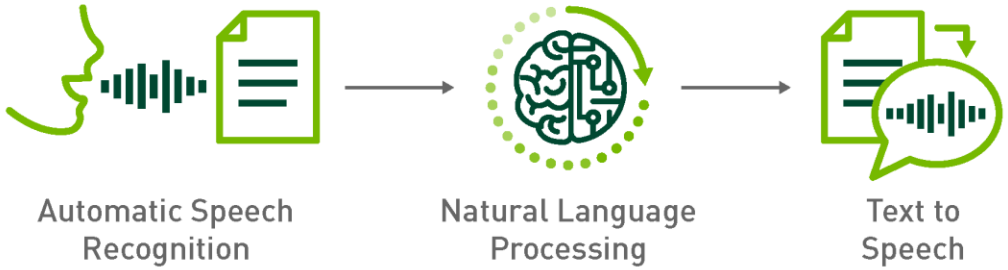
- Definition: Conversational AI that interacts via spoken language
- Uses speech input (microphone) and speech output (speaker)
- Applications: Virtual assistants, customer support bots, smart home devices
- Challenges: Real-time processing, noise robustness, naturalness

GPT4o-Realtime for Voice Agents

- GPT4o-Realtime: Large language model optimized for low-latency streaming
- Enables interactive, natural conversations with voice input/output
- Supports multi-modal (audio + text) integration
- Advantage: Real-time understanding and generation with low delay

Chained Method vs End-to-End Models

- Chained Method:
  - Sequence of models: ASR → LLM → TTS
    - Pros: Modular, easy to debug and improve separately
    - Cons: Accumulated latency, error propagation between modules
- End-to-End Models:
  - Single neural model directly mapping speech to speech or text to speech
    - Pros: Potentially lower latency, holistic optimization
    - Cons: Harder to train, requires massive data



How ASR Works (Automatic Speech Recognition)

- Converts speech audio into text transcription
- Pipeline components:
  - Feature extraction (MFCC, Mel-spectrogram)
  - Acoustic model (neural network, e.g., RNN, Transformer)
  - Language model (contextual understanding)
  - Decoder (beam search to find best transcription)
- Popular ASR models/tools:
  - Whisper (OpenAI)
  - Wav2Vec 2.0 (Facebook)
  - Kaldi (traditional toolkit)

Popular ASR Models and Their Use Cases

Model	Strengths	Use Cases
Whisper	Open-source, multilingual, robust	Podcast transcription, general purpose
Wav2Vec 2.0	Self-supervised pretraining, accurate	Industry-scale ASR, customization
Google Cloud Speech	Cloud-based, scalable	Real-time apps, global languages

How TTS Works (Text-to-Speech)

- Converts text into human-like speech audio
- Pipeline components:
  - Text analysis and normalization
  - Acoustic model generates speech features (mel spectrogram)
  - Vocoder converts features to waveform audio
- Popular TTS architectures:
  - Tacotron 2 + WaveGlow / HiFi-GAN
  - FastSpeech 2
  - VITS (End-to-End TTS)

Popular TTS Models and Their Use Cases

Model	Strengths	Use Cases
Tacotron 2	Natural prosody, widely used	Virtual assistants, audiobooks
FastSpeech 2	Fast inference, controllable prosody	Real-time TTS applications
VITS	End-to-end, high quality	Voice cloning, expressive TTS

Common Audio Data Processing Pipelines

- Typical audio pipeline stages:
  1. Audio capture and noise reduction
  2. Feature extraction (e.g., Mel-spectrogram)
  3. Data augmentation (speed perturbation, noise injection)
  4. Model input preparation

Summary & Key Takeaways

- Voice agents combine ASR, LLM, and TTS for natural spoken interaction
- Modular chained approach is flexible but may add latency
- End-to-end models are promising but resource intensive
- Understanding ASR and TTS pipelines is crucial for building voice agents
- Audio data processing pipelines like Emilia standardize workflows for training

Thank You