Lecture 36: Audio Models Introduction

Introduction:

So far, we have worked with **image models**. Now it's time to explore **audio models** in Spring AI. Audio can be used in two ways:

- **Speech to Text**: Convert spoken audio into written text (Transcription).
- **Text to Speech**: Convert written text into spoken audio.

Types of Audio Models in Spring AI:

- **Transcription (Speech → Text)**
  - Also called **Speech-to-Text (STT)**.
  - You provide an audio file, and the model converts it into text.
  - Common use case: **Captions/Subtitles** for videos.
    - Platforms like **YouTube** and **Udemy** generate captions using transcription models.
  - Useful for accessibility, video content, and automated note generation.

- **Text to Speech (Text → Audio)**
  - Converts a given text into spoken voice output
  - Often called **TTS**.
  - You provide a text input, and the model generates an audio response
  - Options are available to change:
    - **Voice type**
    - **Speed of speech**
  - Use case: Audio learning materials, accessibility tools, or interactive assistants.

Providers for Audio Models:

- **OpenAI**
- **Azure OpenAI** is also supported.

Core Models:

- **OpenAI Transcription Model**
  - Accepts audio input.
  - Uses built-in algorithms to generate accurate text output.
  - Best for caption generation, transcripts, and note-taking.
- **OpenAI Audio Speech Model**
  - Accepts text input.
  - Returns audio output in a natural voice.
  - Provides customization (voice type, speed, etc.).

Key Points:

- **Audio Models** extend AI functionality beyond text and images.
- Two main tasks: **Transcription (STT)** and **Text to Speech (TTS)**.
- Provider: **OpenAI** (primary focus).
- Real-world applications include subtitles, podcasts, accessibility tools, and AI assistants.
- Implementation details will be covered in the next lecture.