Reading: Summary and Highlights

Congratulations! You have completed this module. At this point, you know that:

Multimodal AI

- Processes and understands multiple types of data simultaneously
- Key components: Text processing, computer vision, speech processing, text-to-speech, and multimodal fusion
- Applications: Virtual assistants, healthcare, education, autonomous vehicles, content creation, and accessibility
- Challenges: Data alignment, modality imbalance, resource demands, interpretability, bias in training data, handling sensitive information
- Future trends: Unified models, edge computing, self-supervised learning, personalization, ethical AI, multimodal LLMs

Text-to-Speech (TTS)

- Converts written text into natural-sounding speech
- Combines linguistic analysis with speech synthesis
- Traditional systems: Separate text analysis, acoustic modeling, and waveform generation
- End-to-end systems: Use VAEs, normalizing flows, and GANs (e.g., VITS)
- Applications: Accessibility, virtual assistants, education, entertainment, healthcare, navigation
- Challenges: Natural prosody, emotional context, multiple speakers, real-time optimization, multilingual support

Speech-to-Text (STT)

- Converts spoken language into written text
- Involves audio feature extraction, sound understanding, word prediction, transcription matching, and refinement
- Applications: Captioning, virtual assistants, medical transcription, language learning, note-taking, meeting and court transcription