Automatic Speech Recognition and Beyond

Presenter: Nguyen Manh Tien Anh | Speech Processing Engineer | VinBigdata

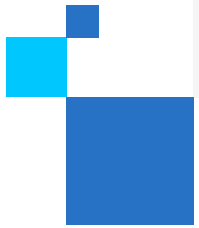
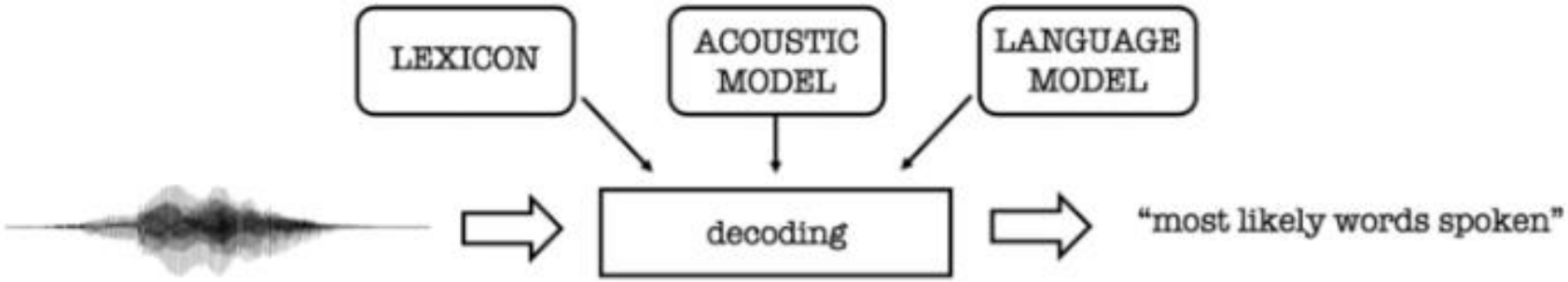


Table of Contents

Overview of ASR

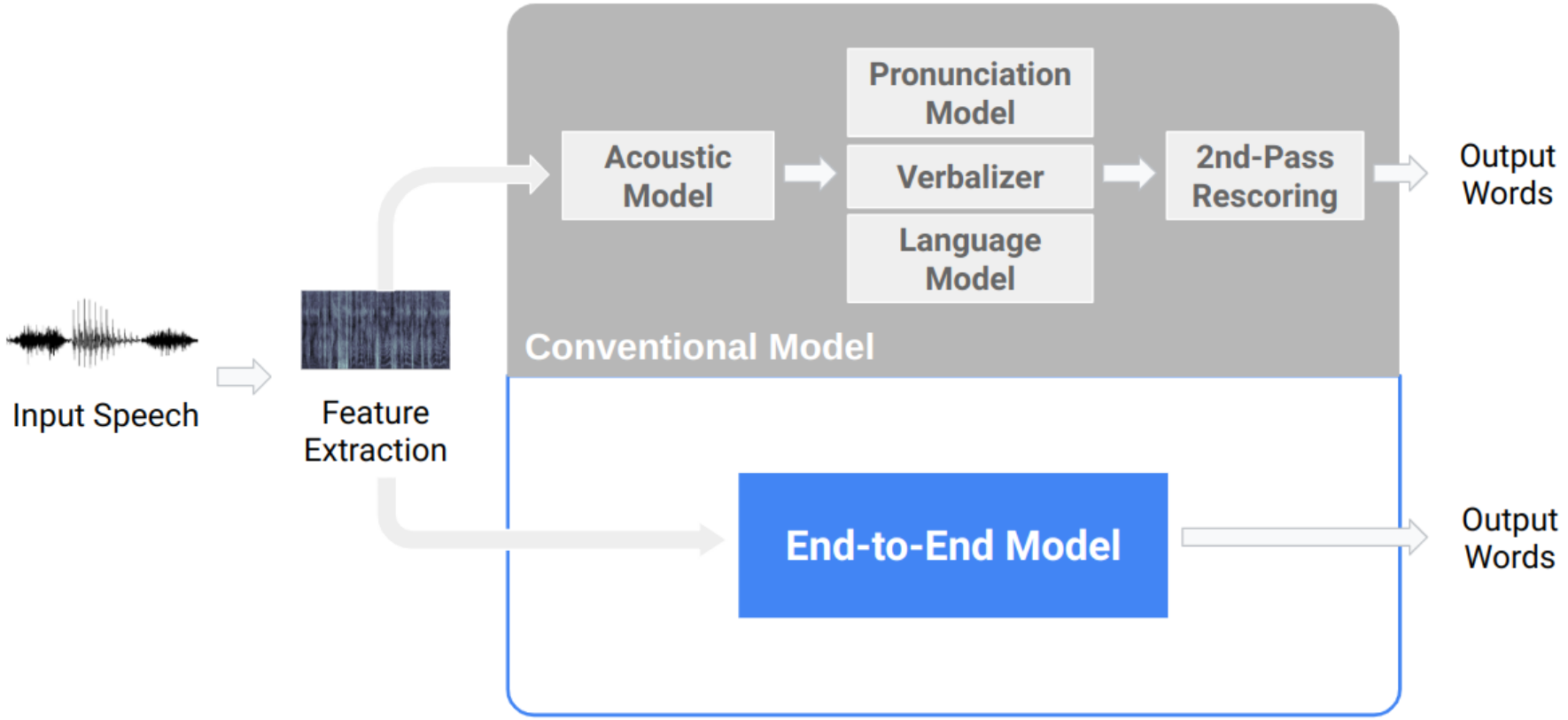
1. Overview of ASR



- ASR aims to convert **continuous audio speech** (input) into **text** (output) with high accuracy, regardless of the **speaker.**
- Main type of ASR:
  - Streaming ASR: Transcribe speech in real-time with low latency.
  - Non-streaming ASR: Transcribe speech after the full audio is received.

1. Overview of ASR

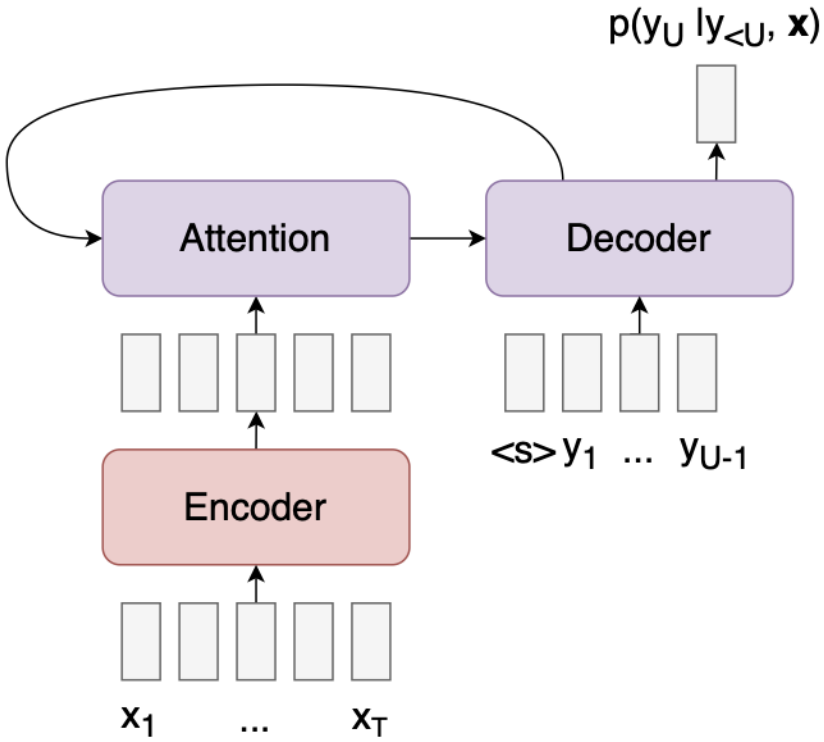
ASR Methodology



1. End-to-end ASR

Attention models

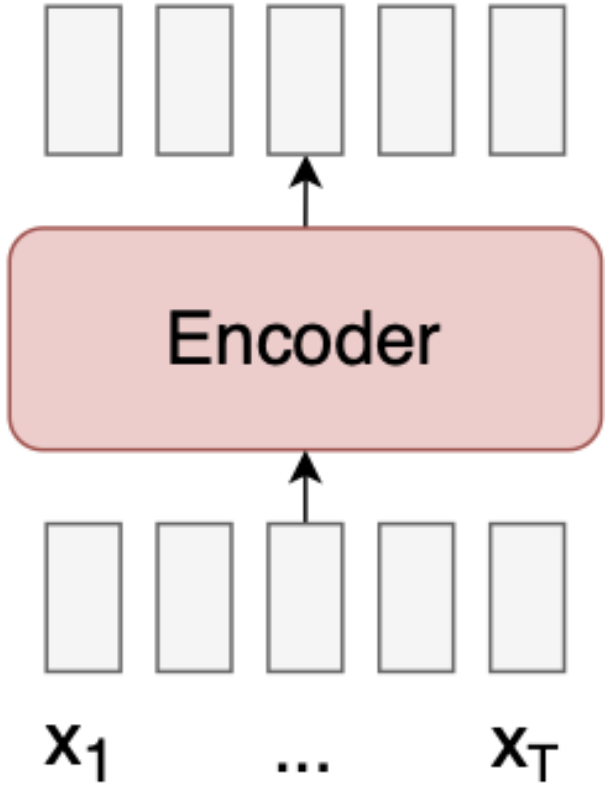
- Intuition: use attention mechanism to focus on different parts of the input for each output
- Advantages:
  - Flexible
  - Captures long-range dependencies
- Disadvantages:
  - Computationally expensive
  - Not suitable for real-time
  - Hard to train for monotonic tasks



1. End-to-end ASR

Connectionist Temporal Classification (CTC) models

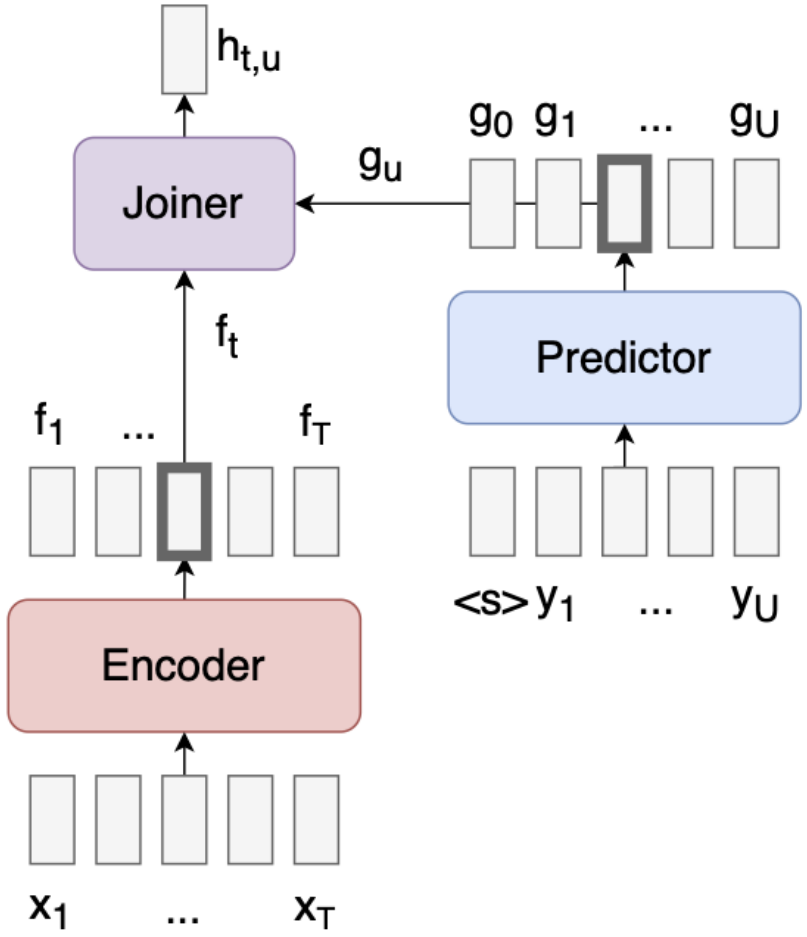
- Intuition: assumes monotonic alignment between input and output => encoder-only architecture
- Advantages:
  - Simple architecture
  - Real-time capacity
- Disadvantages:
  - Outputs are independent
  - Usually use with external Language Model



1. End-to-end ASR

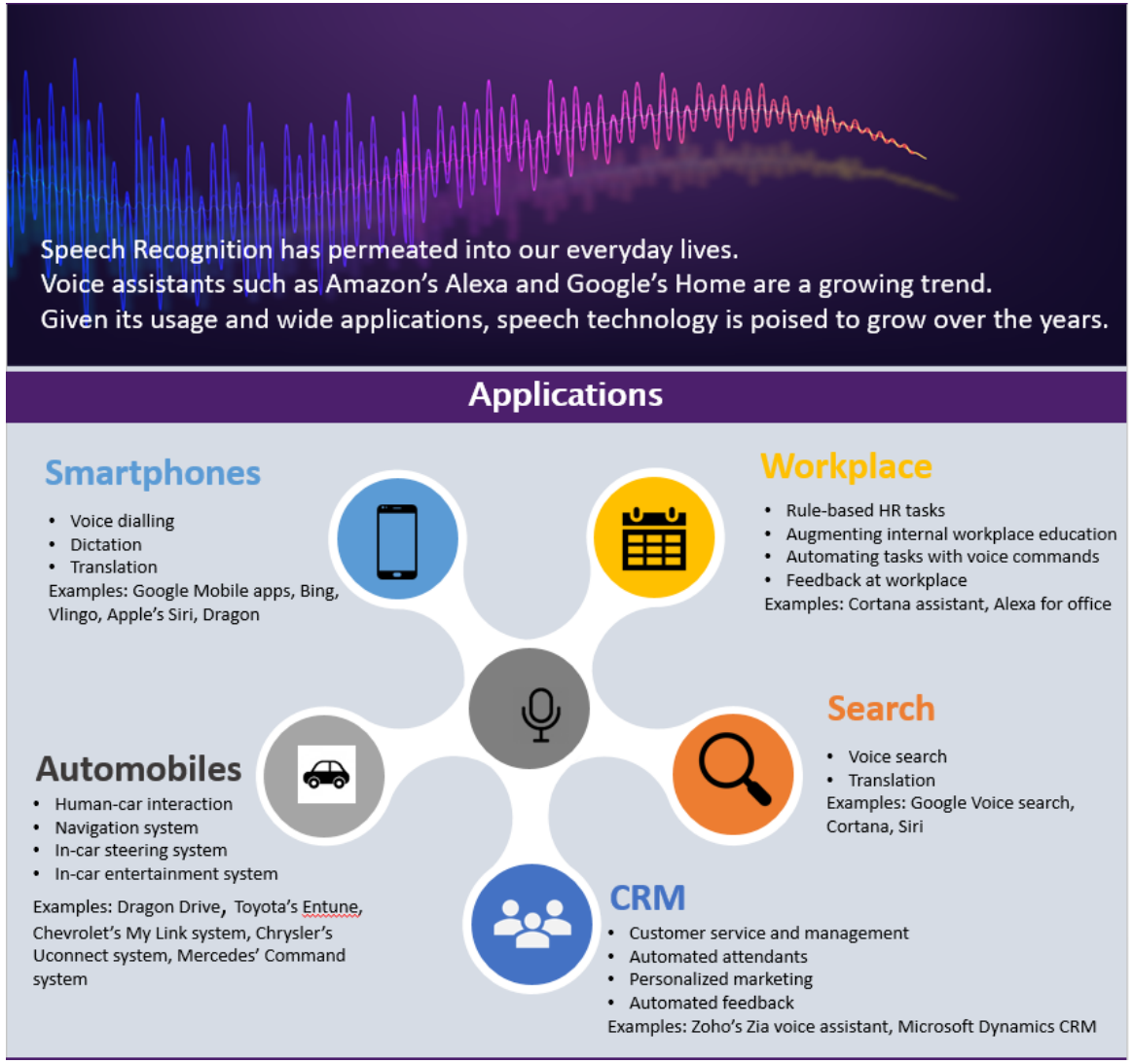
Transducer models

- Intuition: adds a predictor and joiner to handle multiple outputs per input
- Advantages:
  - Handle output dependency
  - Real-time capacity
- Disadvantages:
  - High memory usage
  - Complex training + implementation



1. Overview of ASR

ASR applications

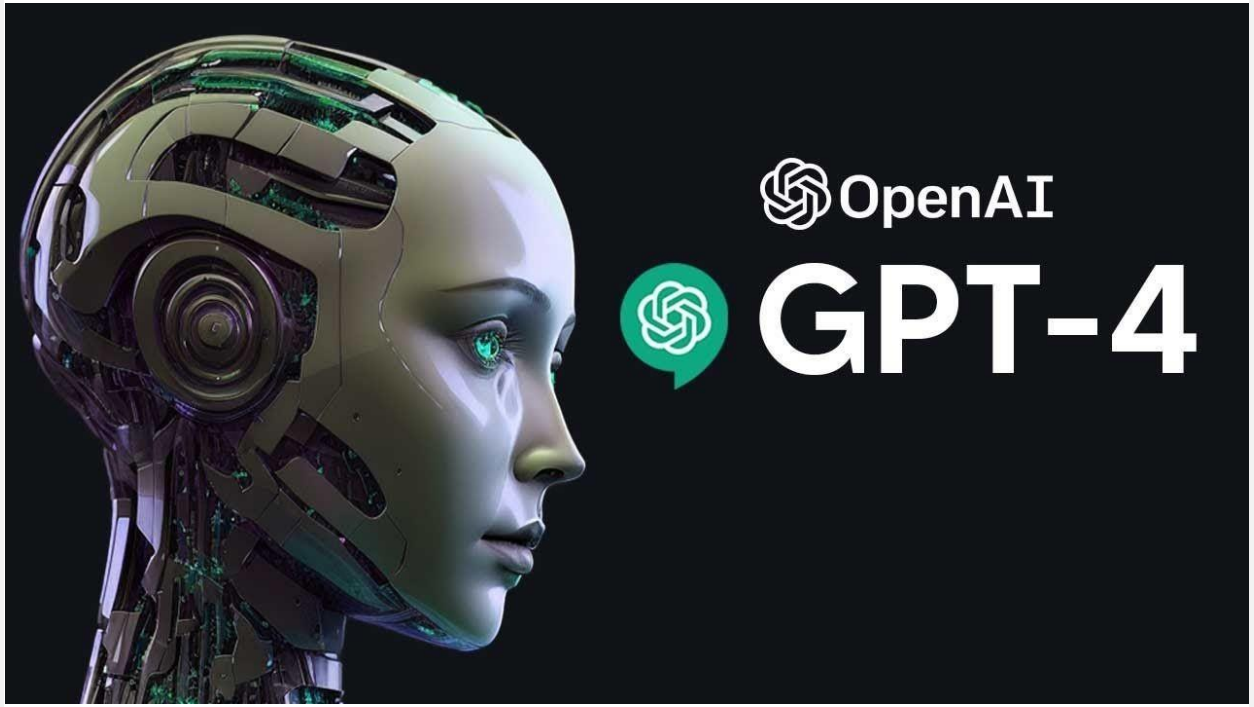


1. Overview of ASR

Challenges

- Diversity: different people speak **differently** depending on their **region**, **age**, **gender**, and **native language**
- **Accents** and **pronunciations** of words
- Environments: Noise, in/out house, …
- **And More**

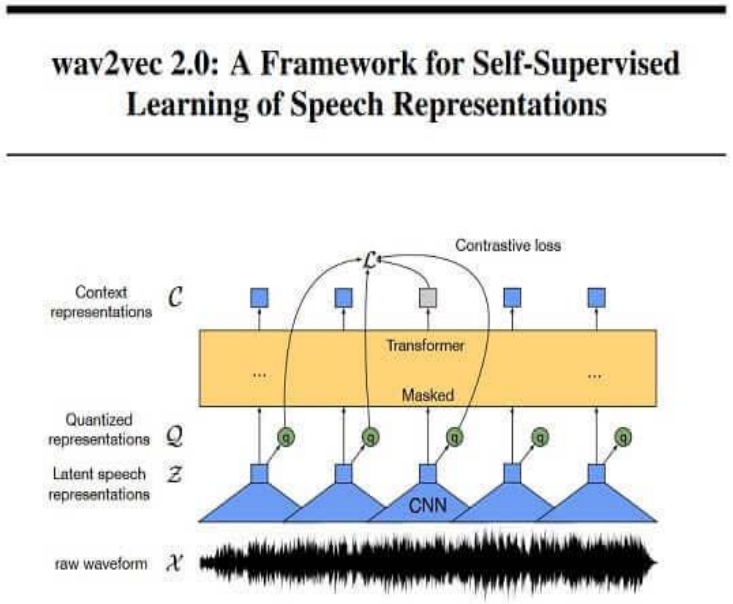
ASR and Beyond



2. ASR and Beyond

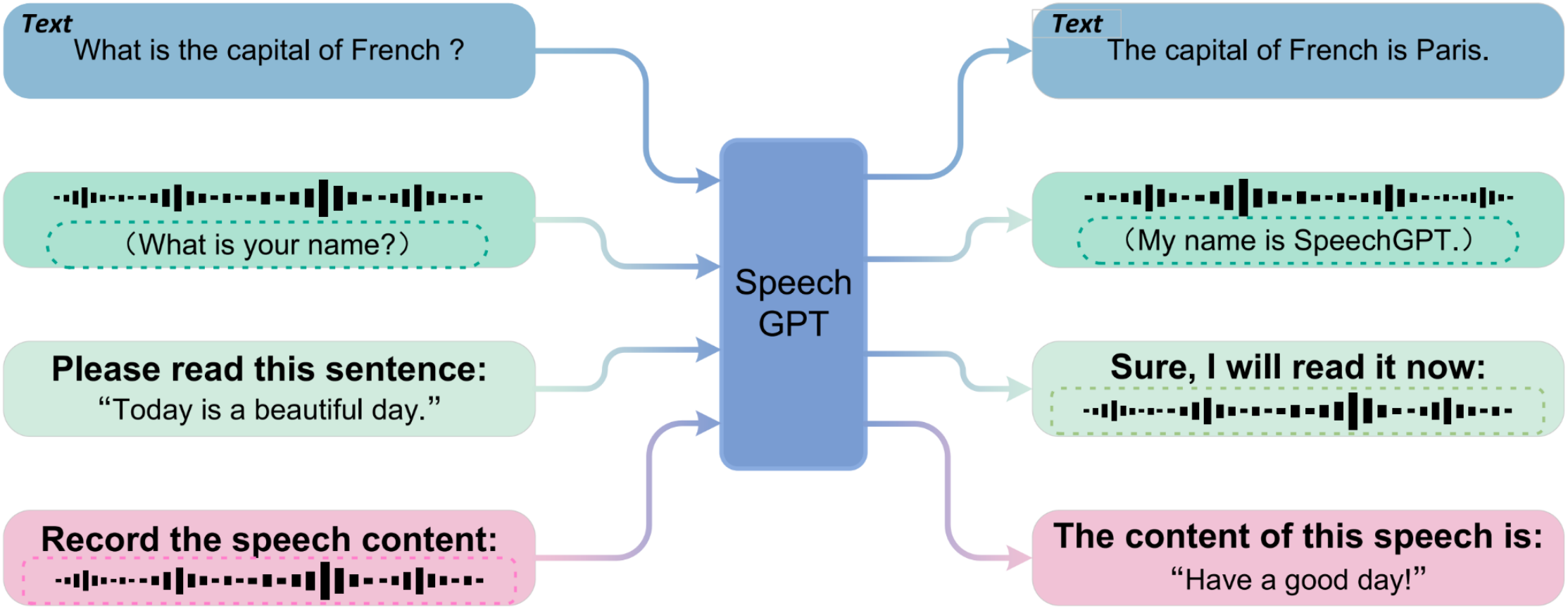
Self-supervised Learning in ASR

- Intuition: learn speech representations from unlabeled data
- SSL tasks: masked prediction or contrastive learning
- Advantages:
  - Efficient use of unlabeled data
  - Better generalization
  - Huge improvement on low-resource language



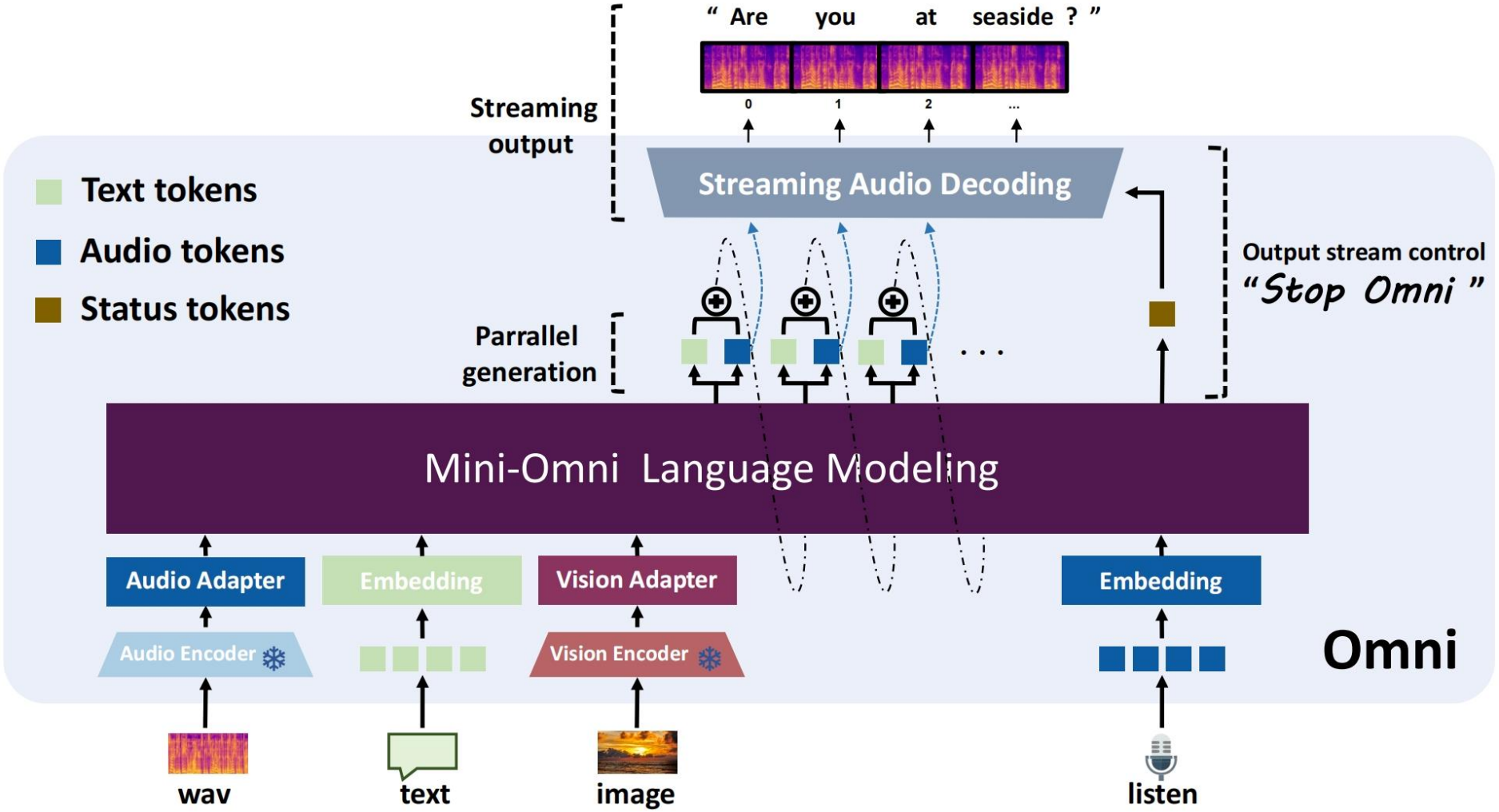
2. ASR and Beyond

LLM + Speech (Multimodal models)



2. ASR and Beyond

LLM + Speech + Image (Multimodal models)



3. Speech Processing Division / ASR Team

Welcome to ASR Team!

- Many interesting and complex problems – R&D:
  - SSL, Multilingual, Streaming ASR,...
  - Domain Adaptation, Semantic Endpoint,...
  - Optimize model: CUDA/Ondevice…
  - Ondevice ASR, Cloud ASR,...
  - Multimodal
- Improve software engineering skills
- Competitions/Scientific Conference
- Great working environment
- Great supervisors
- More…

Thank you for listening!