



Recent Advances in End-to-End Automatic Speech Recognition

Jinyu Li

Outline

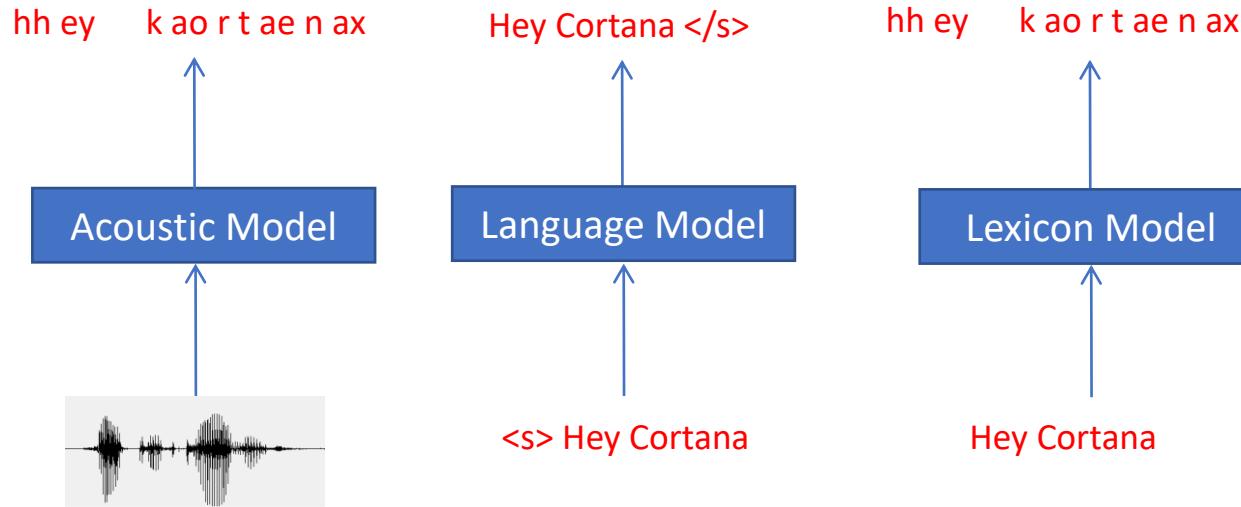
- End-to-end (E2E) automatic speech recognition (ASR) fundamental
- E2E advances
 - Leveraging unpaired text
 - Multilingual ASR
 - Multi-talker ASR
 - Beyond ASR
- The next trend
- Conclusions

End-to-End Fundamental

Hybrid vs. End-to-End (E2E) Modeling

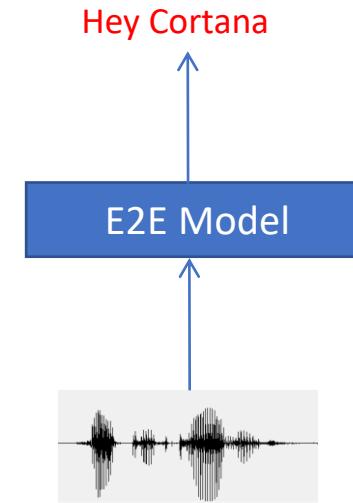
Hybrid

Separate models are trained, and then are used all together during testing in an ad-hoc way.



E2E

A single model is used to directly map the speech waveform into the target word sequence.



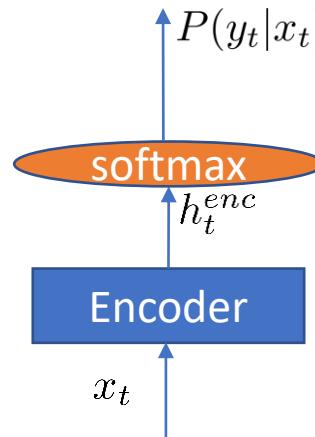
Advantages of E2E Models

- E2E models use a single objective function which is consistent with the ASR objective
- E2E models directly output characters or even words, greatly simplifying the ASR pipeline
- E2E models are much more compact than traditional hybrid models -- can be deployed to devices with high accuracy and low latency

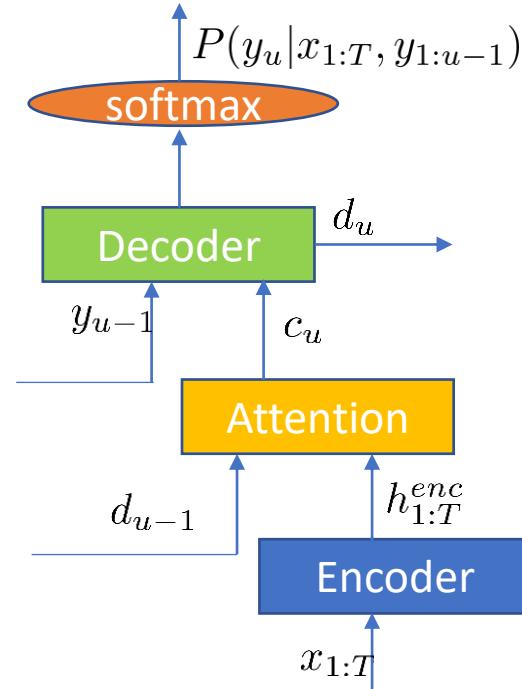
Current Status

- E2E models achieve the state-of-the-art results in most benchmarks in terms of ASR accuracy.
- Practical challenges such as streaming, latency, adaptation capability etc., have been also optimized in E2E models.
- E2E models are now the mainstream models not only in academic but also in industry.

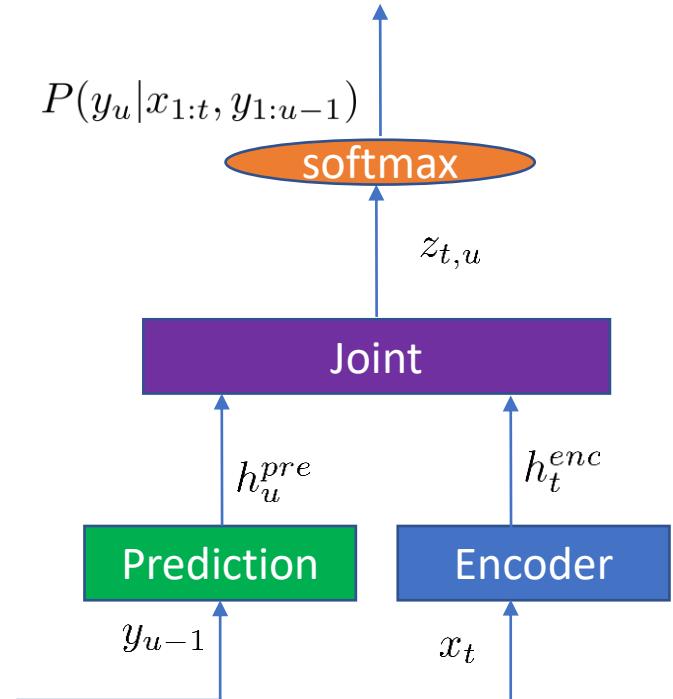
E2E Models



Connectionist Temporal Classification (CTC)



Attention-based encoder decoder (AED)



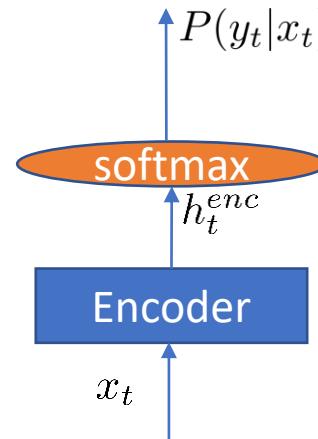
RNN-Transducer (RNN-T)

E2E Models

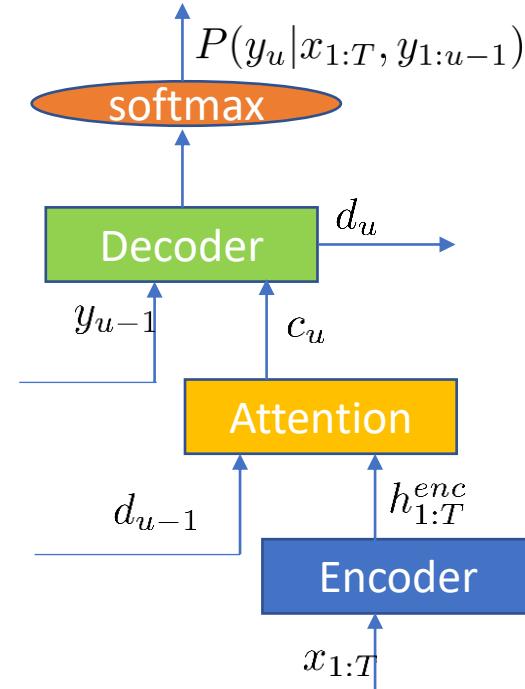
	CTC	AED	RNN-T
Independence assumption	Yes	No	No
Attention mechanism	No	Yes	No
Streaming	Natural	Additional work needed	Natural
Ideal operation scenario	Streaming	Offline	Streaming
Long form capability	Good	Weak	Good

RNN-T is the most popular E2E model in industry which requires streaming ASR.

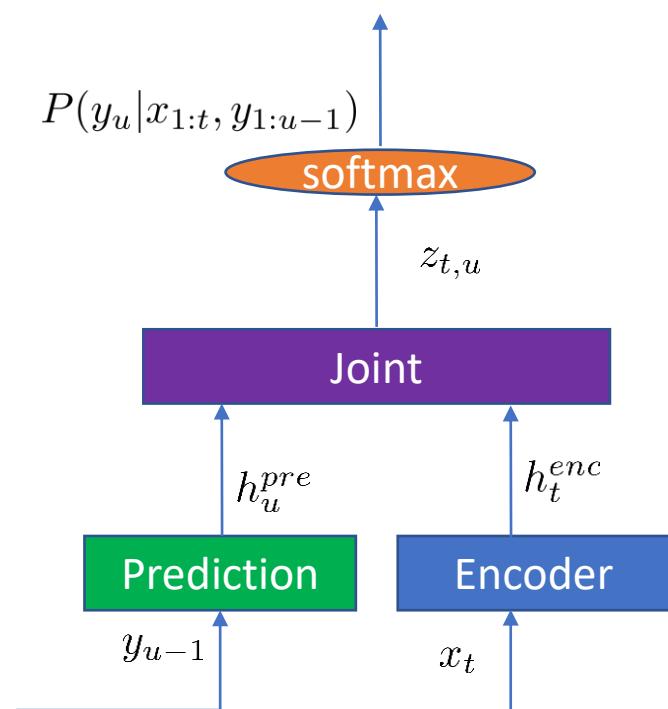
Encoder is the Most Important Component



Connectionist Temporal
Classification (CTC)

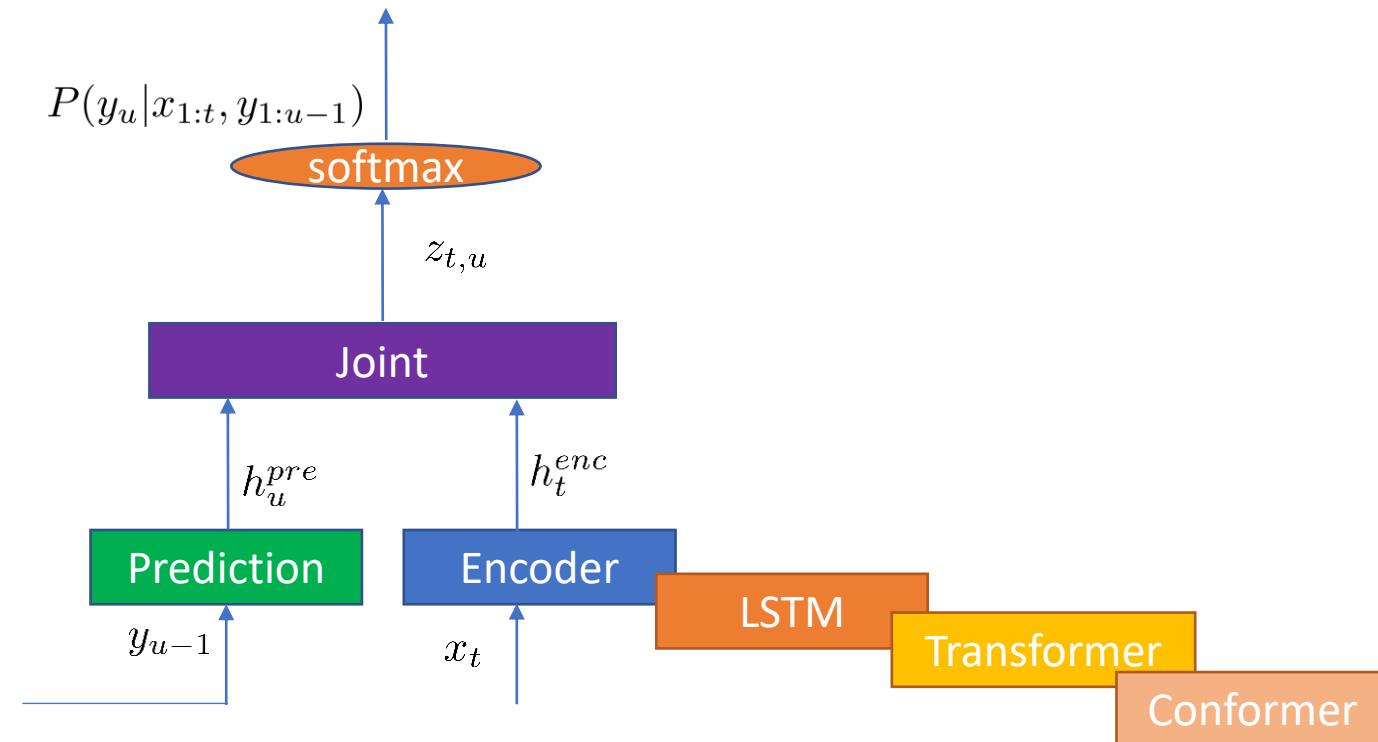


Attention-based encoder
decoder (AED)



RNN-Transducer
(RNN-T)

Encoder for RNN-T



Transformer

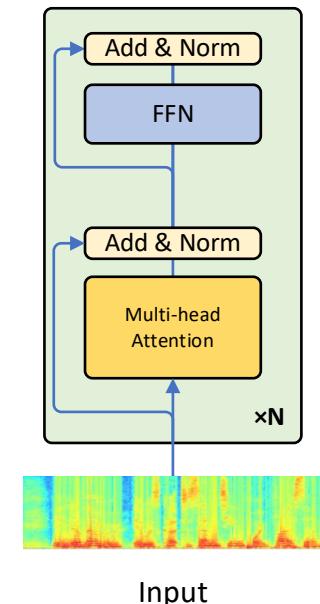
- Self-attention: computes the attention distribution over the input speech sequence

$$\alpha_{t,\tau} = \frac{\exp(\beta(\mathbf{W}_q \mathbf{x}_t)^T (\mathbf{W}_k \mathbf{x}_\tau))}{\sum_{\tau'} \exp(\beta(\mathbf{W}_q \mathbf{x}_t)^T (\mathbf{W}_k \mathbf{x}_{\tau'}))}$$

- Attention weights are used to combine the value vectors to generate the layer output

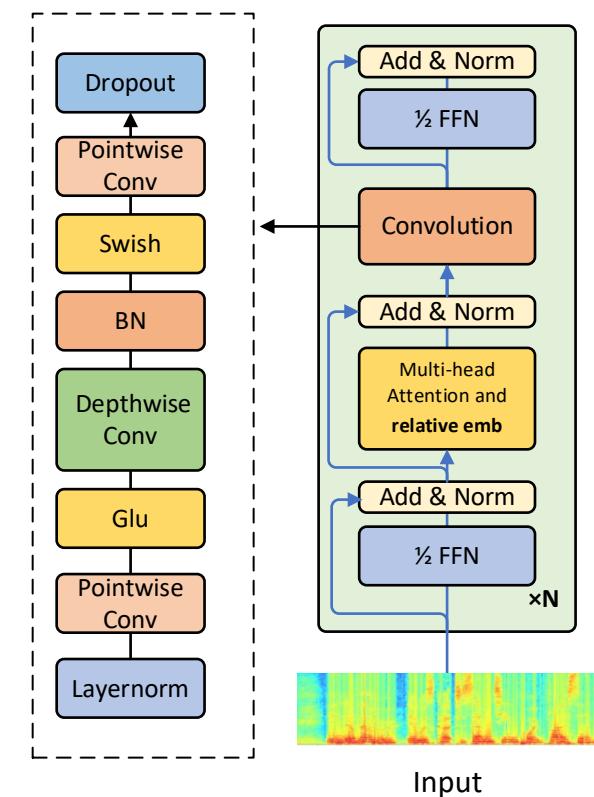
$$\mathbf{z}_t = \sum_{\tau} \alpha_{t\tau} \mathbf{W}_v \mathbf{x}_\tau = \sum_{\tau} \alpha_{t\tau} \mathbf{v}_\tau$$

- Multi-head self-attention: applies multiple parallel self-attentions on the input sequence



Conformer

- Transformer: good at capturing global context, but less effective in extracting local patterns
- Convolutional neural network (CNN): works on local information
- Conformer: combines Transformer with CNN



Gulati et al. "Conformer: Convolution-augmented Transformer for Speech Recognition," in Proc. Interspeech, 2020.

Industry Requirement of Transformer Encoder

- Streaming with low latency and low computational cost
- Vanilla Transformer fails so because it attends the full sequence
- Solution: Attention mask is all you need

Attention Mask is All You Need

- Compute attention weight $\{\alpha_{t,\tau}\}$ for time t over input sequence $\{x_\tau\}$,
binary attention mask $\{m_{t,\tau}\}$ to control range of input $\{x_\tau\}$ to use

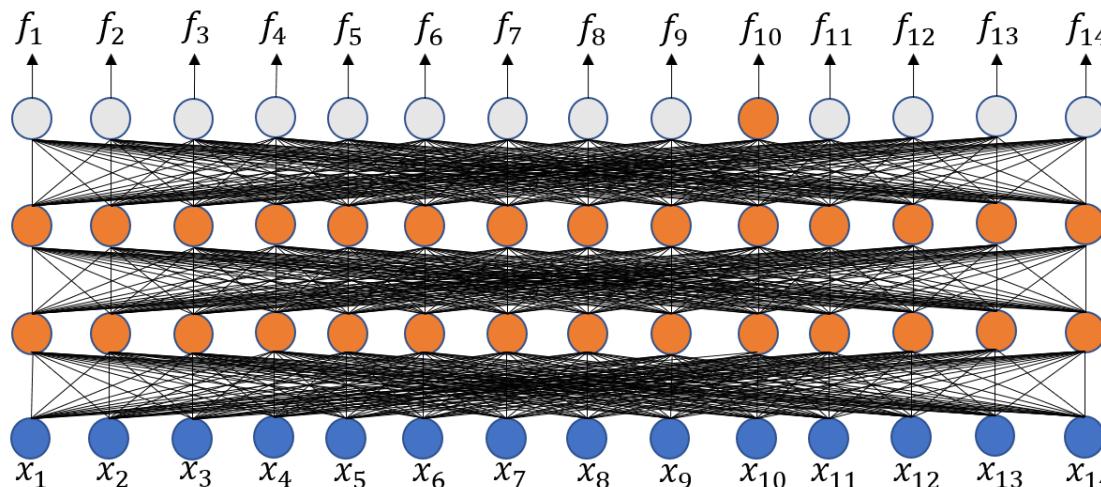
$$\alpha_{t,\tau} = \frac{m_{t,\tau} \exp(\beta (W_q x_t)^T (W_k x_\tau))}{\sum_{\tau'} m_{t,\tau'} \exp(\beta (W_q x_t)^T (W_k x_{\tau'}))} = \text{softmax}(\beta q_t^T k_\tau, m_{t,\tau})$$

- Apply attention weight over value vector $\{v_\tau\}$

$$z_t = \sum_{\tau} \alpha_{t,\tau} W_v x_\tau = \sum_{\tau} \alpha_{t,\tau} v_\tau$$

Attention Mask is All You Need

- Offline (whole utterance)



Predicting output for x_{10}

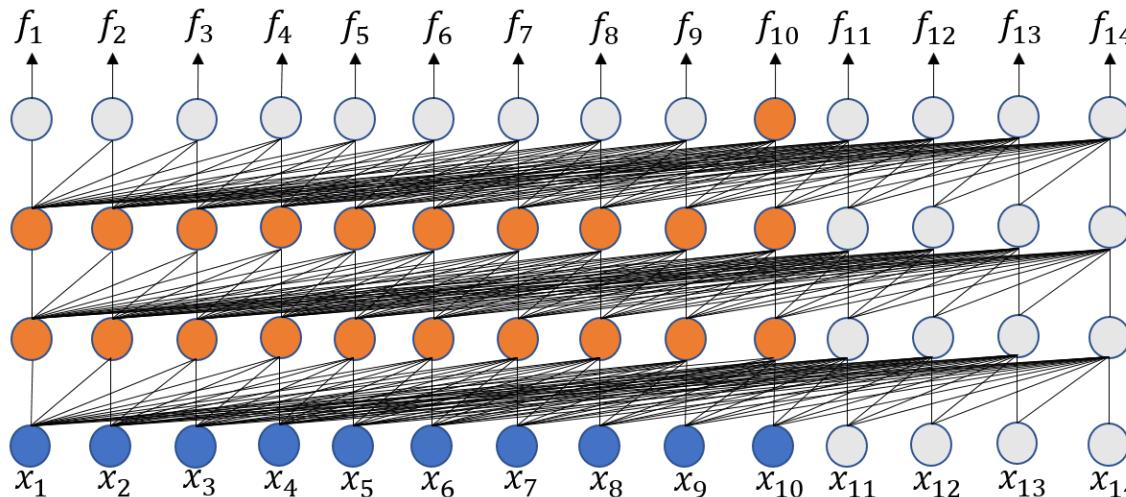
Not streamable

Frame Index	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1

Attention Mask

Attention Mask is All You Need

- 0 lookahead, full history



Predicting output for x_{10}

**Memory and runtime cost
increase linearly**

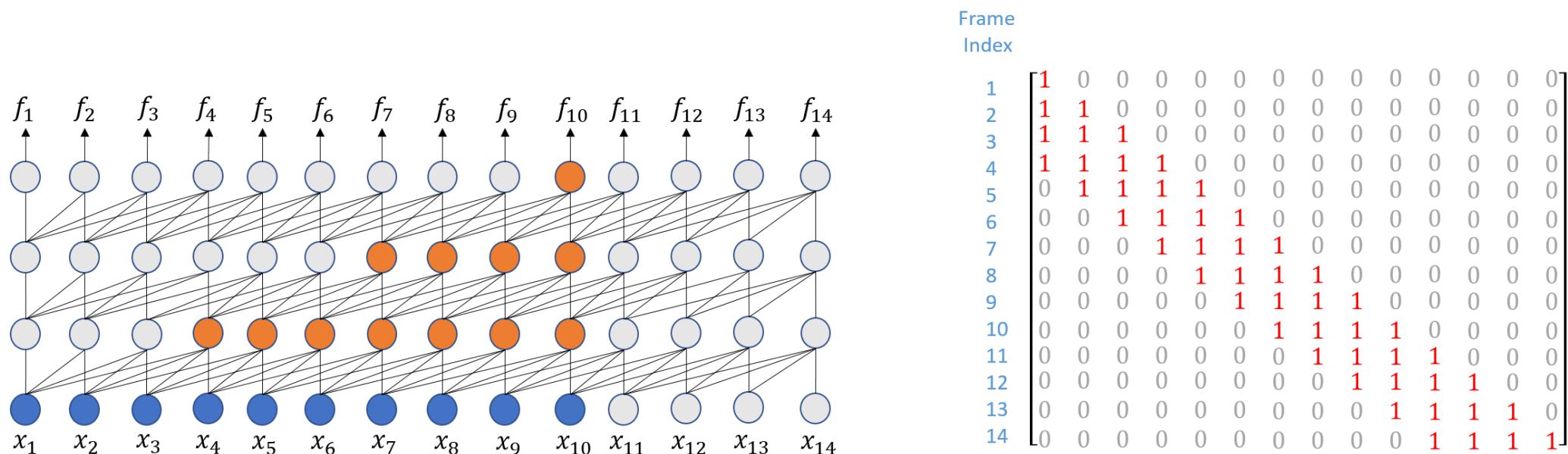
Frame Index

Frame Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	1	0	0	0	0
11	1	1	1	1	1	1	1	1	1	1	1	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	1	0	0
13	1	1	1	1	1	1	1	1	1	1	1	1	1	0
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Attention Mask

Attention Mask is All You Need

- 0 lookahead, limited history (3 frames)



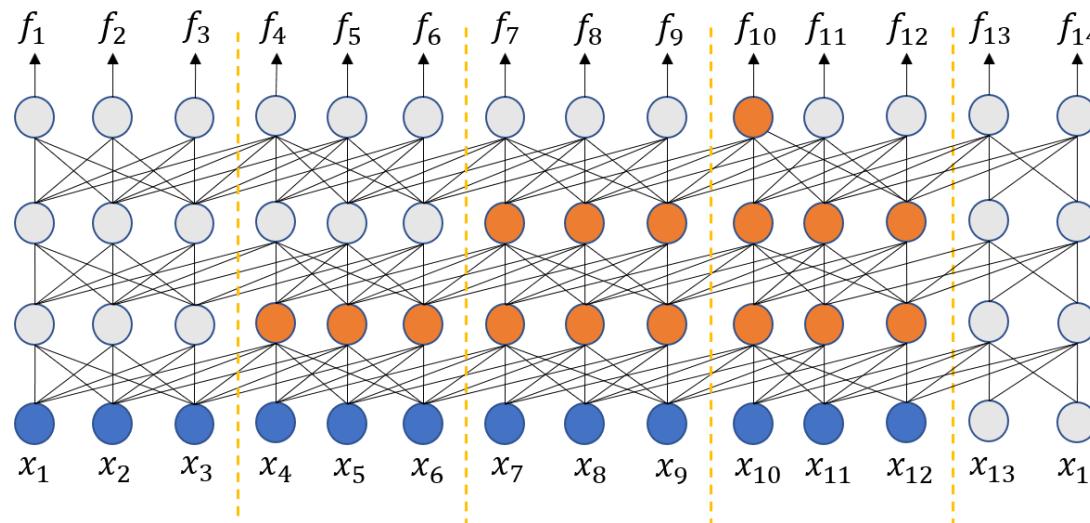
Predicting output for x_{10}

**In some scenario, small amount
of latency is allowed**

Attention Mask

Attention Mask is All You Need

- Small lookahead (at most 2 frames), limited history (3 frames)



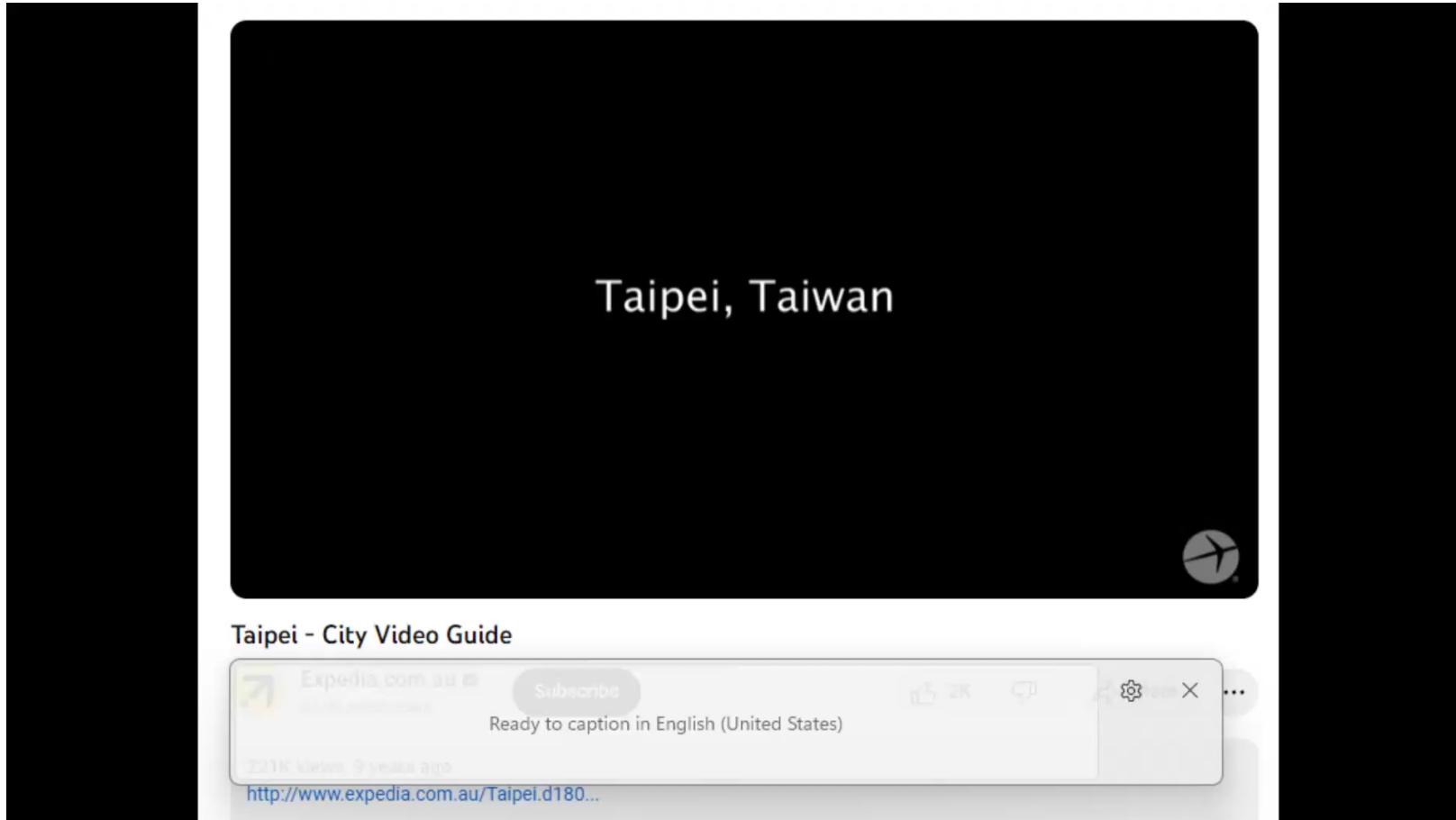
Look-ahead window [0, 2]

Predicting output for x_{10}

Frame Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	0	0	0	0	0	0	0	0
7	0	0	0	1	1	1	1	1	1	1	1	0	0	0
8	0	0	0	1	1	1	1	1	1	1	1	0	0	0
9	0	0	0	1	1	1	1	1	1	1	1	0	0	0
10	0	0	0	0	0	0	1	1	1	1	1	1	1	0
11	0	0	0	0	0	0	1	1	1	1	1	1	1	0
12	0	0	0	0	0	0	1	1	1	1	1	1	1	0
13	0	0	0	0	0	0	0	0	0	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Attention Mask

Live Caption in Windows 11



Advancing E2E Models

unpaired
text

multi-talker

multilingual

speech
translation

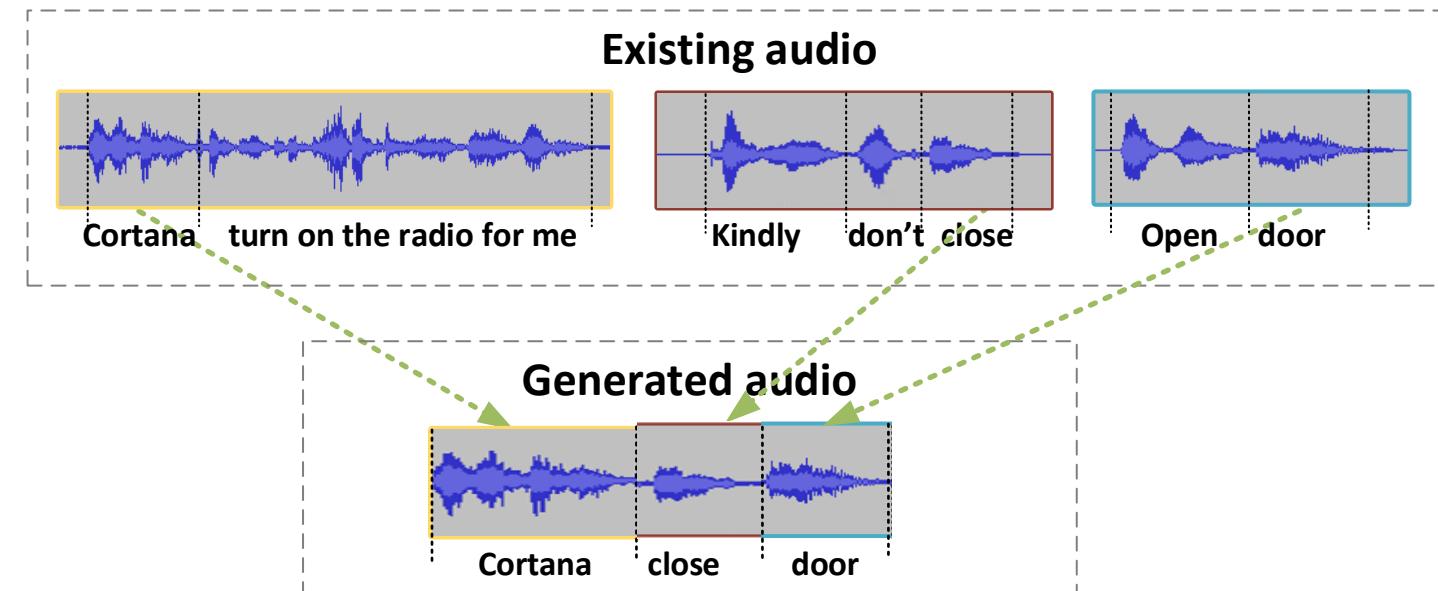
Unpaired Text

Leverage Unpaired Text

- Standard E2E models are trained with paired speech-text data, while hybrid models use large amount of text data for LM building.
- It is important to leverage unpaired text data for further performance improvement, especially in the domain adaptation task.
 - Adaptation with augmented audio
 - LM fusion
 - Direct adaptation with text data

Adaptation with Augmented Audio

- Adapt E2E models with the synthesized speech generated from the new domain text using TTS or from original ASR training data.



Sim, K., et al. Personalization of end-to-end speech recognition on mobile devices for named entities. *in Proc. ASRU*, 2019.

Li, J., et al. Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability. *in Proc. Interspeech*, 2020.

Zheng, X., et al., Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems. *in Proc. ICASSP*, 2021.

Zhao, R., et al., On Addressing Practical Challenges for RNN-Transducer. *in Proc. ASRU*, 2021.

LM Fusion Methods

- Shallow Fusion
 - A log-linear interpolation between the E2E and LM probabilities.

$$\hat{Y} = \operatorname{argmax}_Y \left[\log P(Y|X; \theta_{\text{E2E}}^S) + \lambda_T \log P(Y; \theta_{\text{LM}}^T) \right]$$

E2E score
Target LM score

- Density Ratio Method
 - Subtract source-domain LM score from Shallow Fusion score.

$$\hat{Y} = \operatorname{argmax}_Y \left[\log P(Y|X; \theta_{\text{E2E}}^S) + \lambda_T \log P(Y; \theta_{\text{LM}}^T) - \lambda_S \log P(Y; \theta_{\text{LM}}^S) \right]$$

Shallow Fusion score
Source LM score

A standalone LM trained with training transcript of E2E model

- HAT/ILME-based Fusion
 - Subtract internal LM score from Shallow Fusion score.
 - Show improved ASR performance over Shallow Fusion and Density Ratio

$$\hat{Y} = \operatorname{argmax}_Y \left[\log P(Y|X; \theta_{\text{E2E}}^S) + \lambda_T \log P(Y; \theta_{\text{LM}}^T) - \lambda_I \log P(Y; \theta_{\text{E2E}}^I) \right]$$

Shallow Fusion score
Internal LM score

An inherent LM estimated by E2E model parameters

Gulcehre, C., et al. On using monolingual corpora in neural machine translation. arXiv:1503.03535, 2015.

McDermott, E., et al. A density ratio approach to language model fusion in end-to-end automatic speech recognition. in *Proc. ASRU*, 2019.

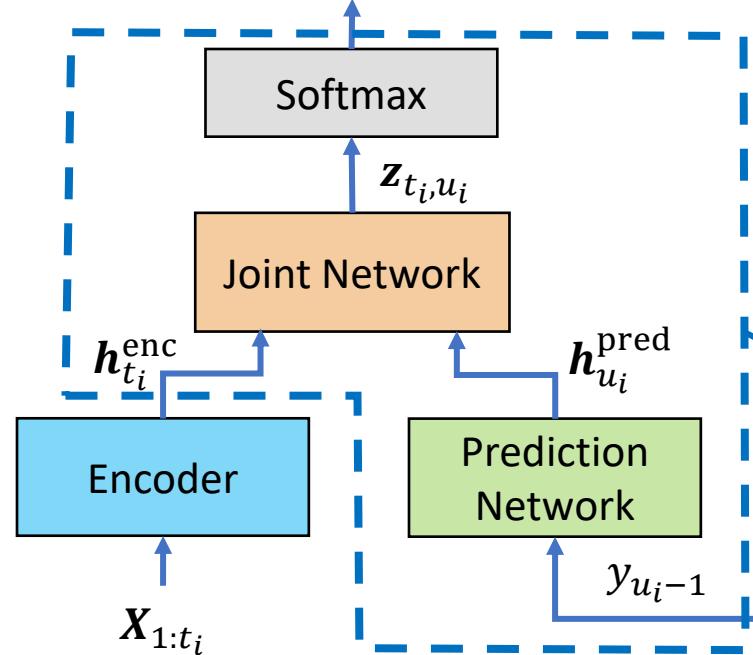
Variani, E., et al. Hybrid autoregressive transducer (HAT). in *Proc. ICASSP*, 2020.

Meng, Z., et al. Internal language model estimation for domain-adaptive end-to-end speech recognition. in *Proc. SLT*, 2021.

Internal LM Estimation

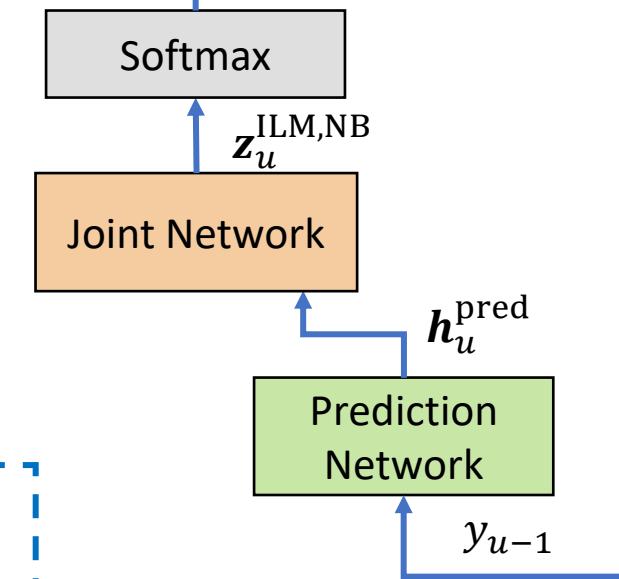
► RNN-T

$$P(\tilde{y}_i | \mathbf{Y}_{0:u_i-1}, \mathbf{X}_{1:t_i}; \theta_{\text{RNN-T}}) = \text{softmax}(\mathbf{z}_{t_i, u_i})$$



► Internal LM estimation of RNN-T

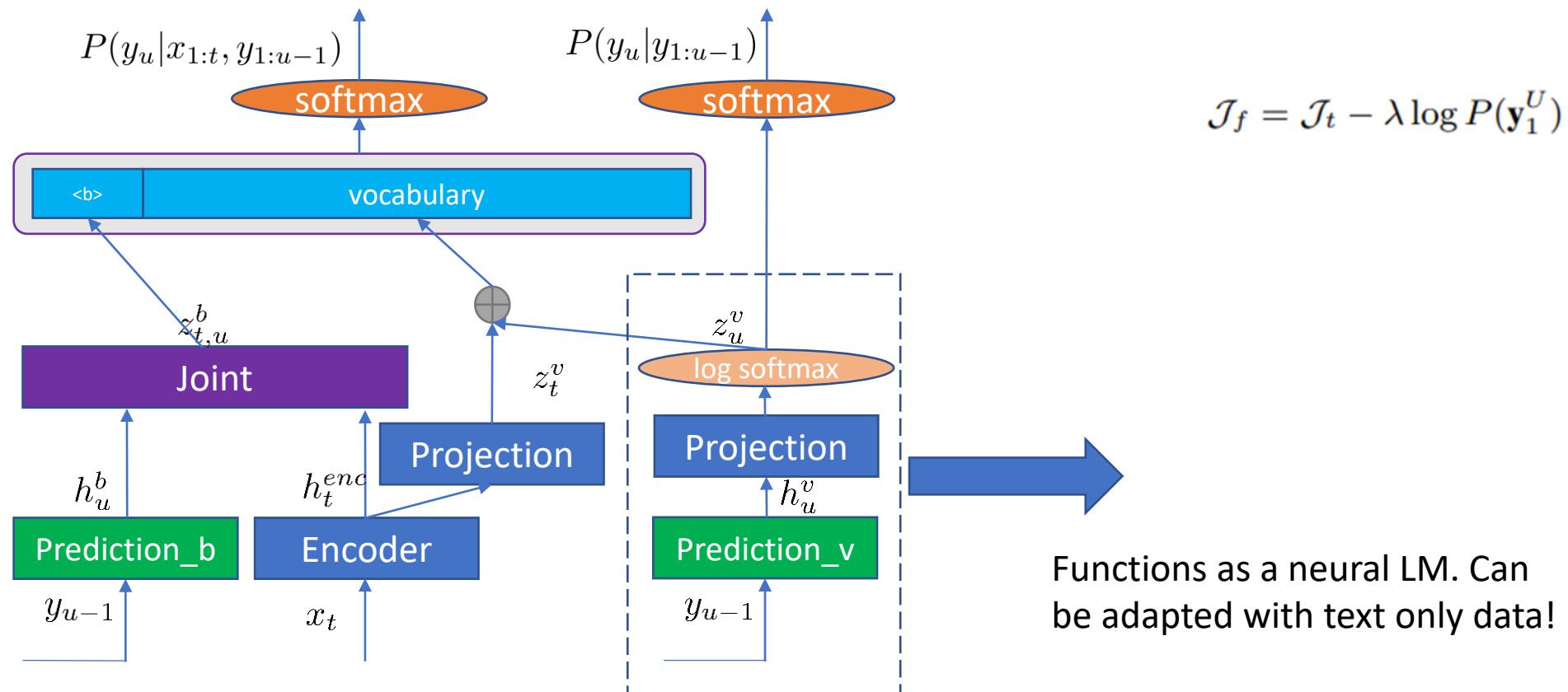
$$P(y_u | \mathbf{Y}_{0:u-1}; \theta_{\text{pred}}, \theta_{\text{joint}}) = \text{softmax}(\mathbf{z}_u^{\text{ILM,NB}})$$



- Internal LM probability

➤ The output of the **acoustically-conditioned LM** after removing the contribution of the encoder

Factorized Neural Transducer



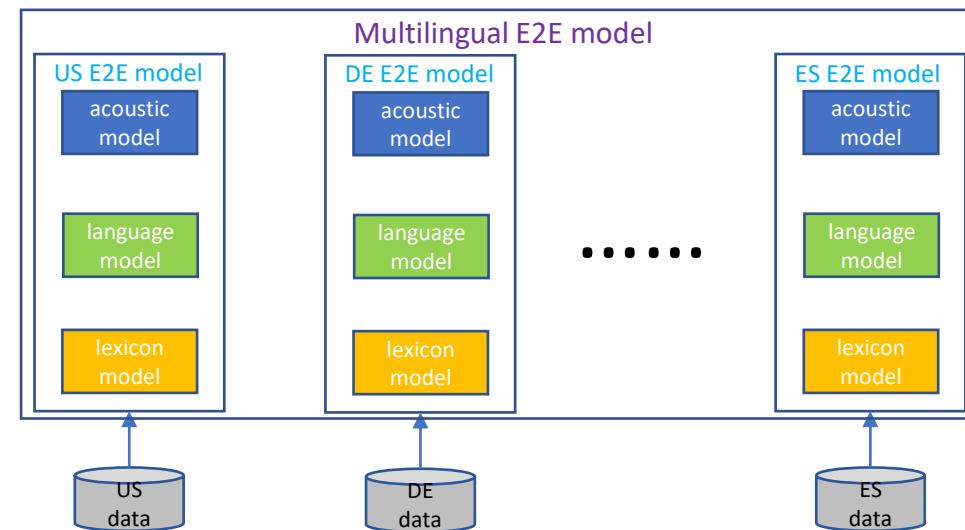
Multilingual ASR

Multilingual

- 40% people can speak only 1 language fluently.
- 43% people can speak only 2 languages fluently.
- 13% people can speak only 3 languages fluently.
- 3% people can speak only 4 languages fluently.
- <0.1% people can speak 5+ languages fluently.
- Human cannot recognize all languages. Can we build a *single high quality multilingual model* to serve *all users*?

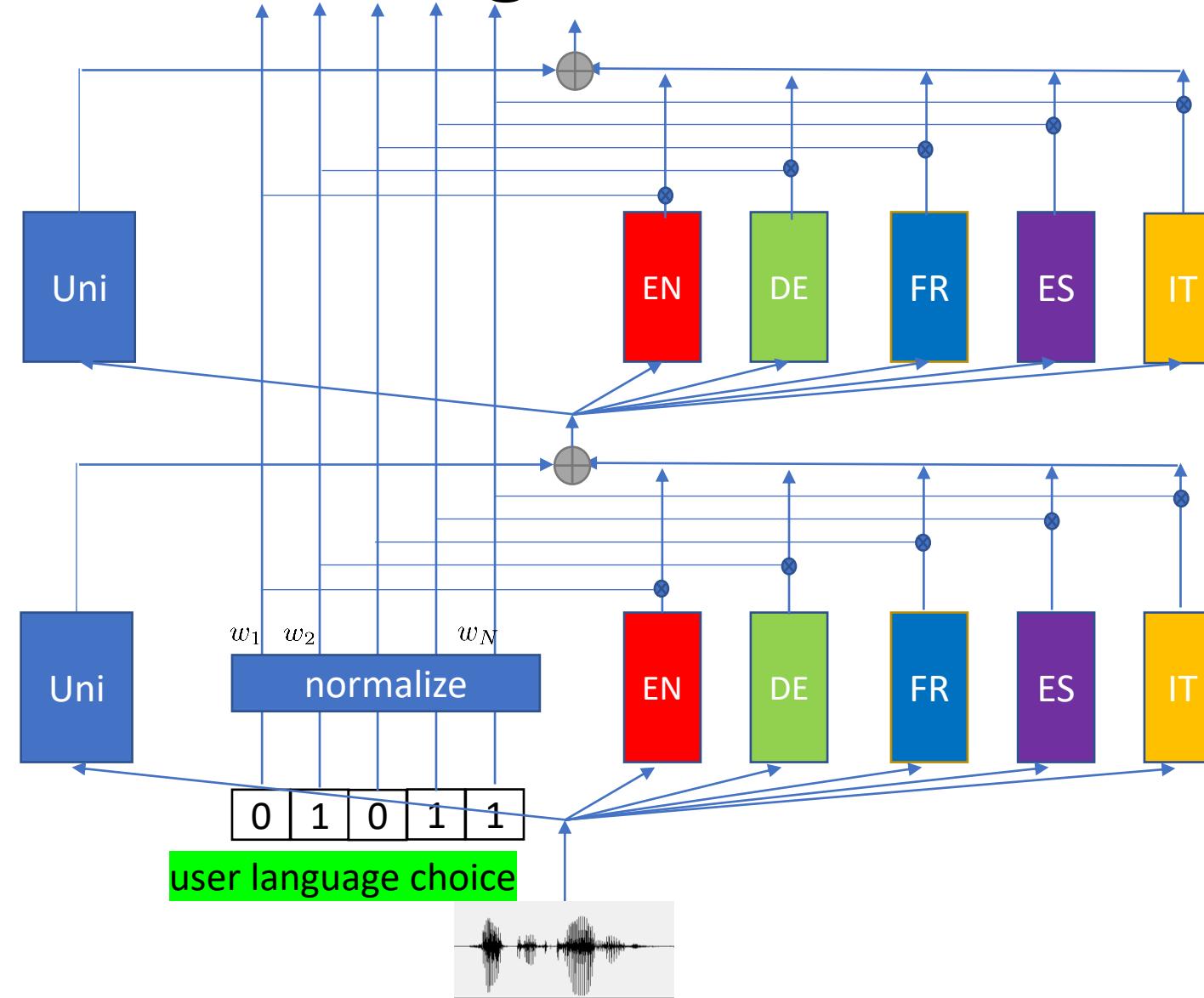
Multilingual E2E Models

- Double-edged sword of pooling all language data
 - Maximum sharing between languages; One model for all languages
 - Confusion between languages



Configurable Multilingual Model

- **Universal module:** modeling the sharing across languages
- **Expert module:** modeling the residual from universal module for each language



Multi-talker ASR

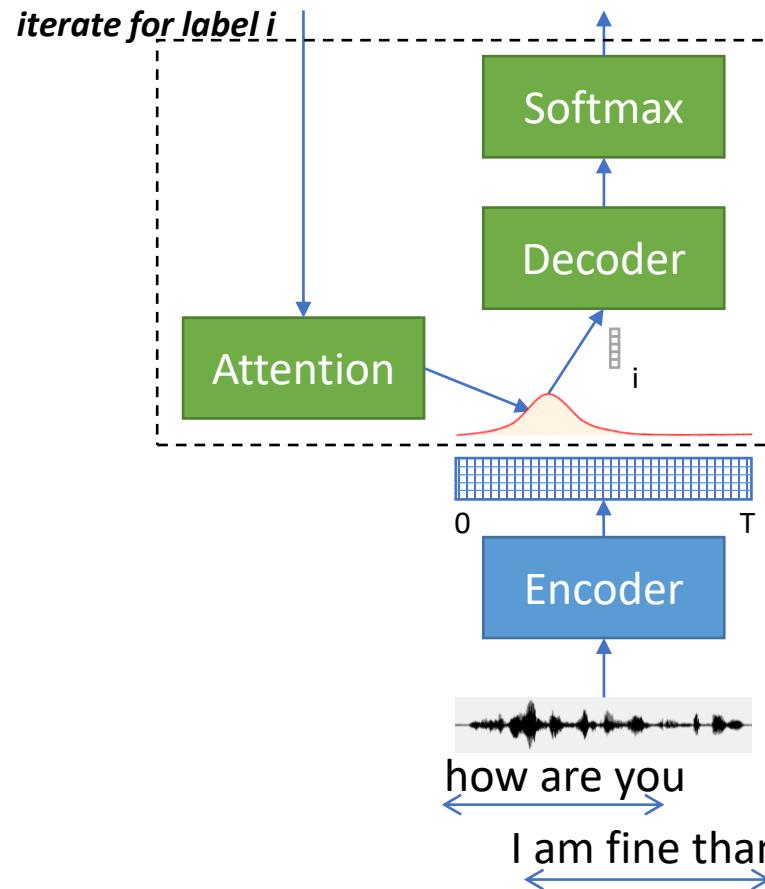
Multi-talker ASR

- E2E ASR systems have high accuracy in single-speaker applications 😊
- Very difficult to achieve satisfactory accuracy in scenarios with multiple speakers talking at the same time 😞
- Solutions: E2E multi-talker models

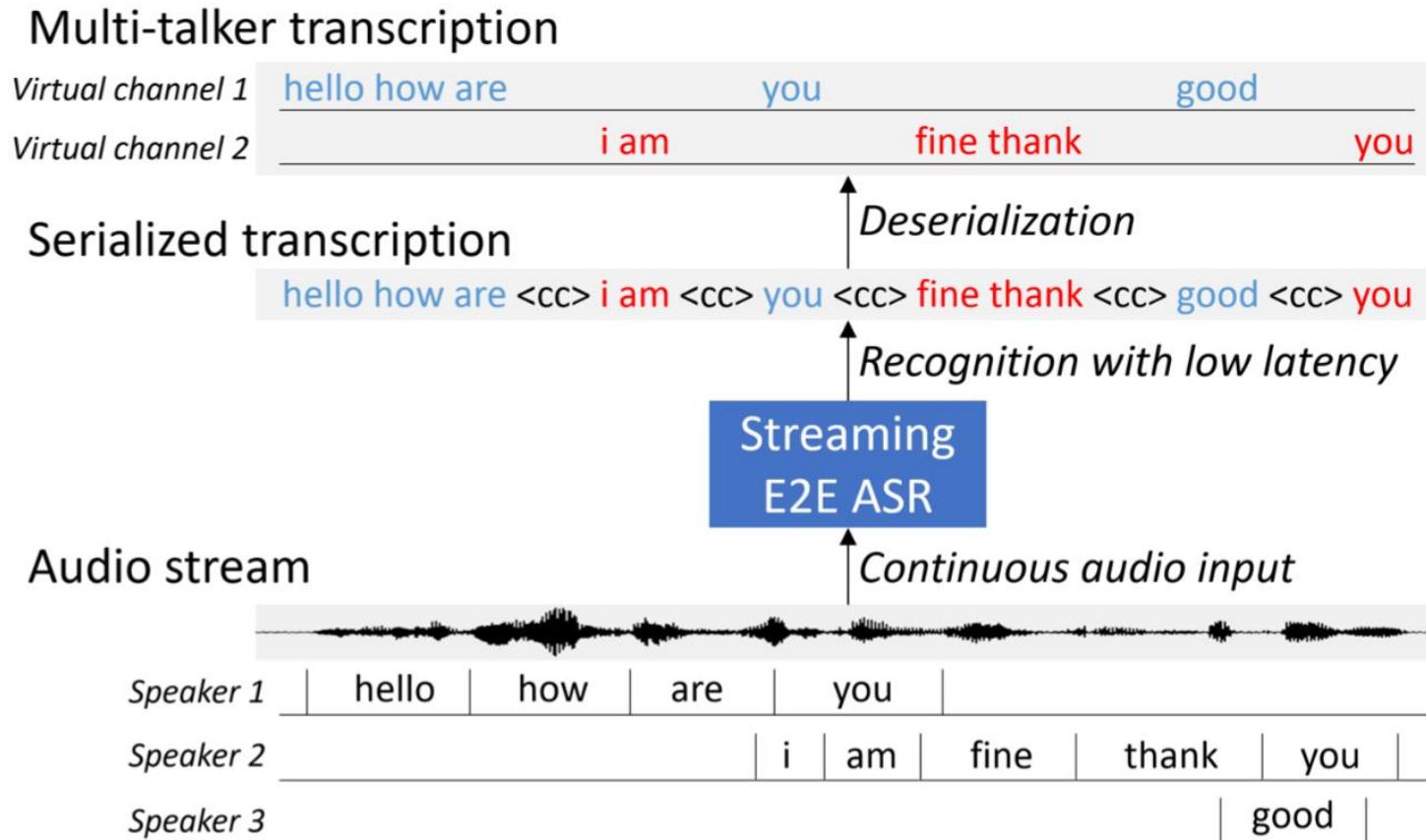
Serialized Output Training (SOT)



how are you <sc> I am fine thank you <eos>



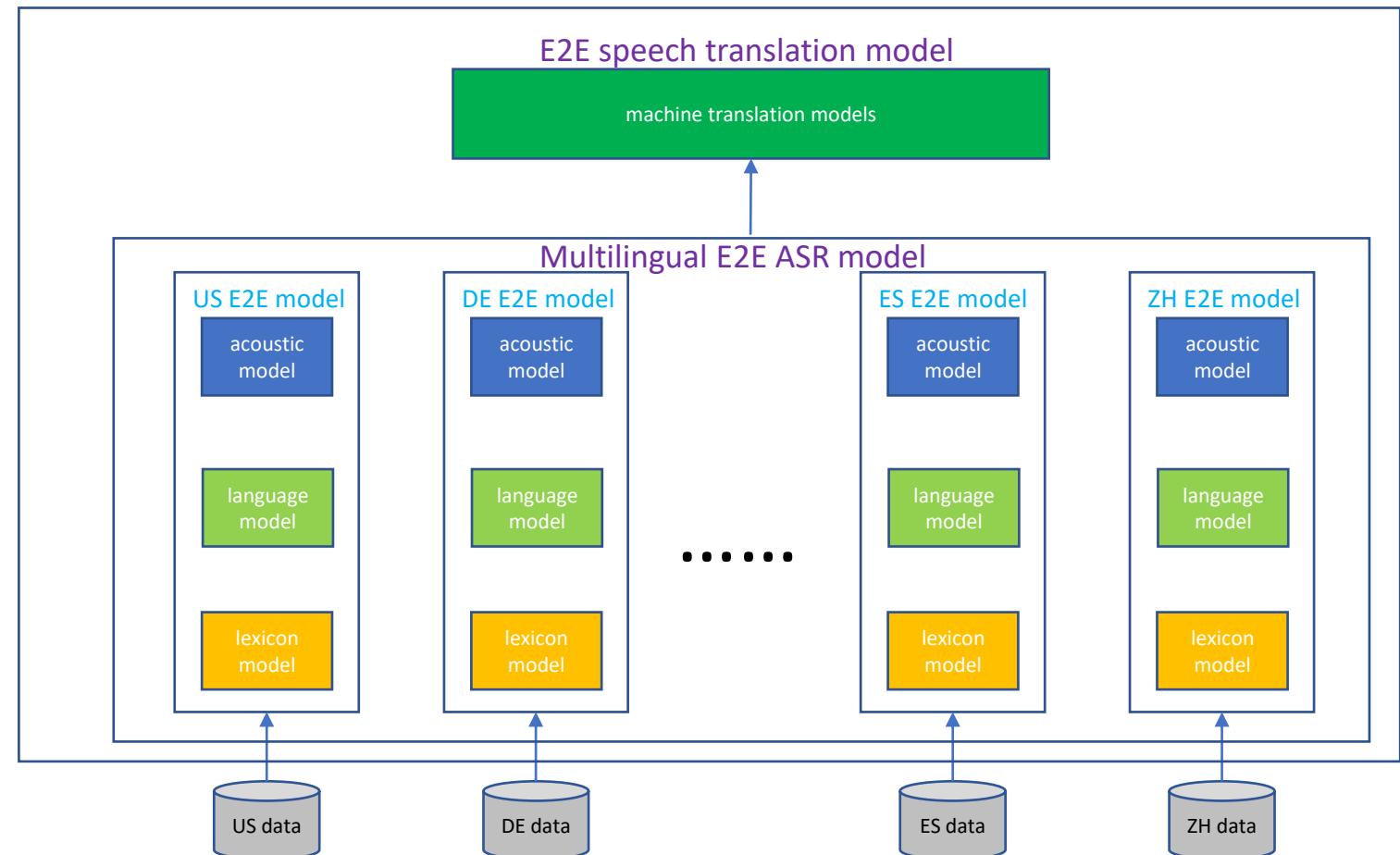
Token-level Serialized Output Training (t-SOT)



Beyond ASR

E2E Speech Translation (ST)

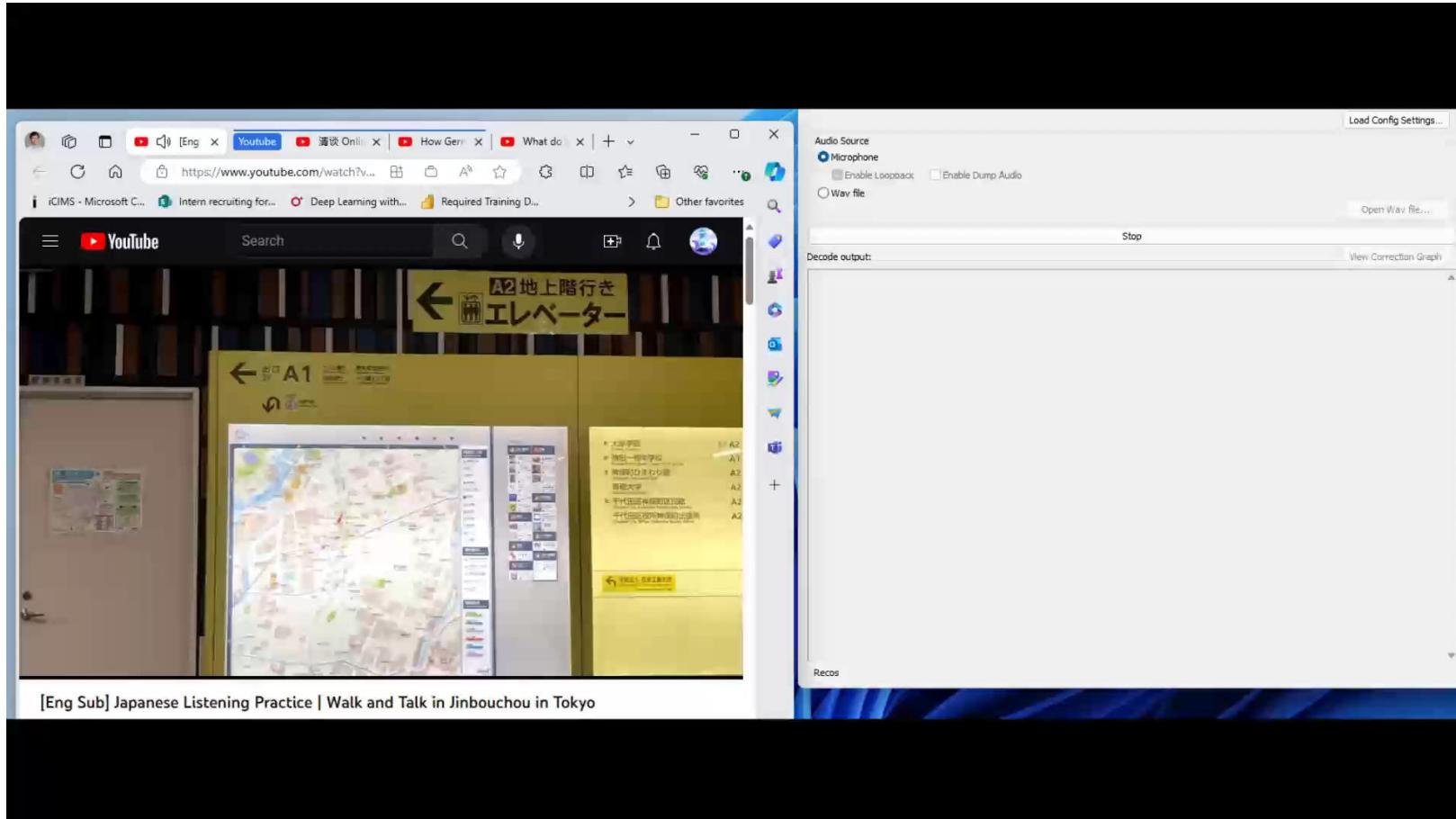
- ASR is often the first step in a pipeline and is followed by
 - machine translation
 - speech synthesis (→ speech-to-speech translation)
 - natural language understanding / generation, etc.



Streaming Multilingual Speech Model (SM²)

- Multilingual data is pooled together to train a streaming model to perform both ST and ASR functions.
- ST training is totally weakly supervised without using any human labeled parallel corpus.
- The model is very small, running on devices.

Simultaneous ST Demo

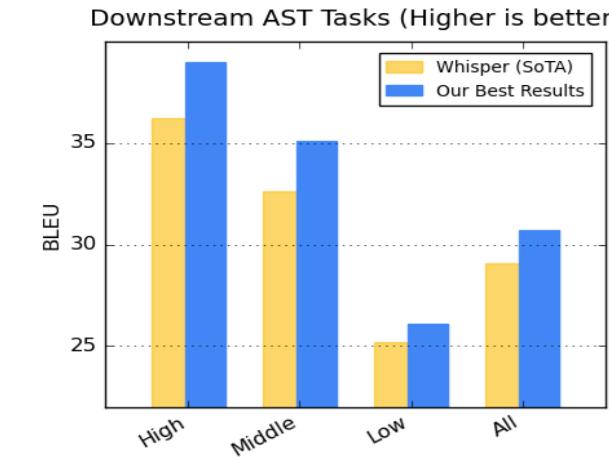
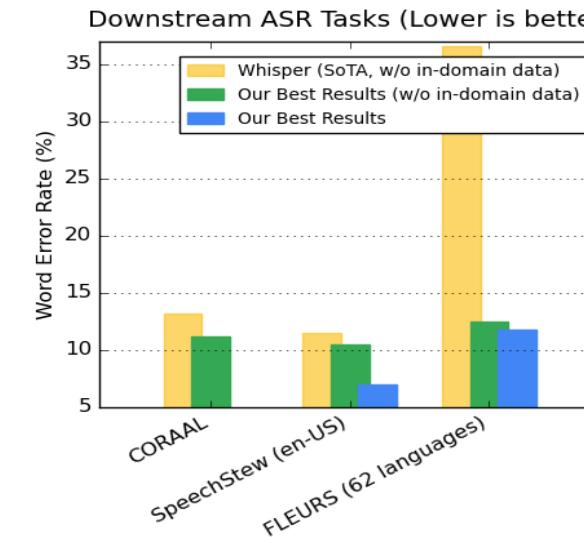
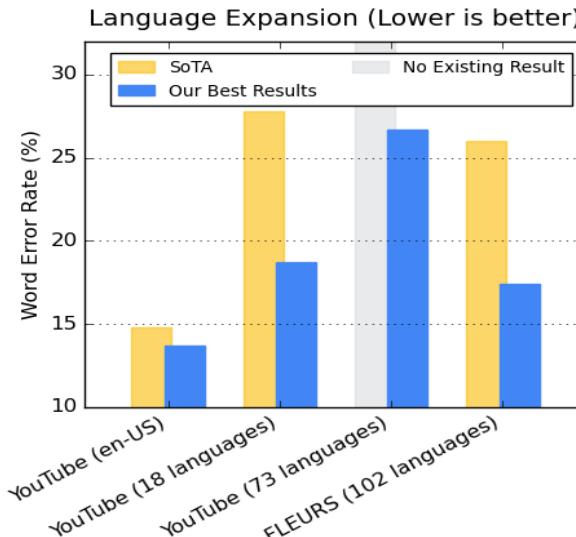
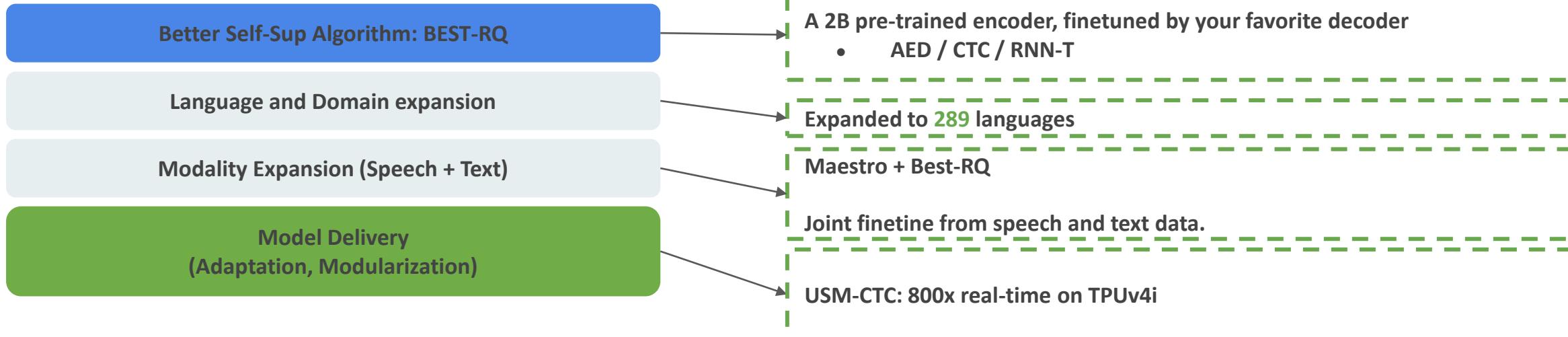


Foundation Models

Whisper from OpenAI

- Trained from 680k hours weakly supervised data collected from the web.
- A single model can perform multiple tasks: multilingual ASR + speech translation (to English), language identification, etc.
- Outstanding zero-shot capability

Universal Speech Understanding (USM) model



Credit to Yu Zhang

Zhang, Y., et al. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, arXiv:2303.01037, 2023.

What's the Next Trend?

Large Language Model (LLM)-centric Integrative AI

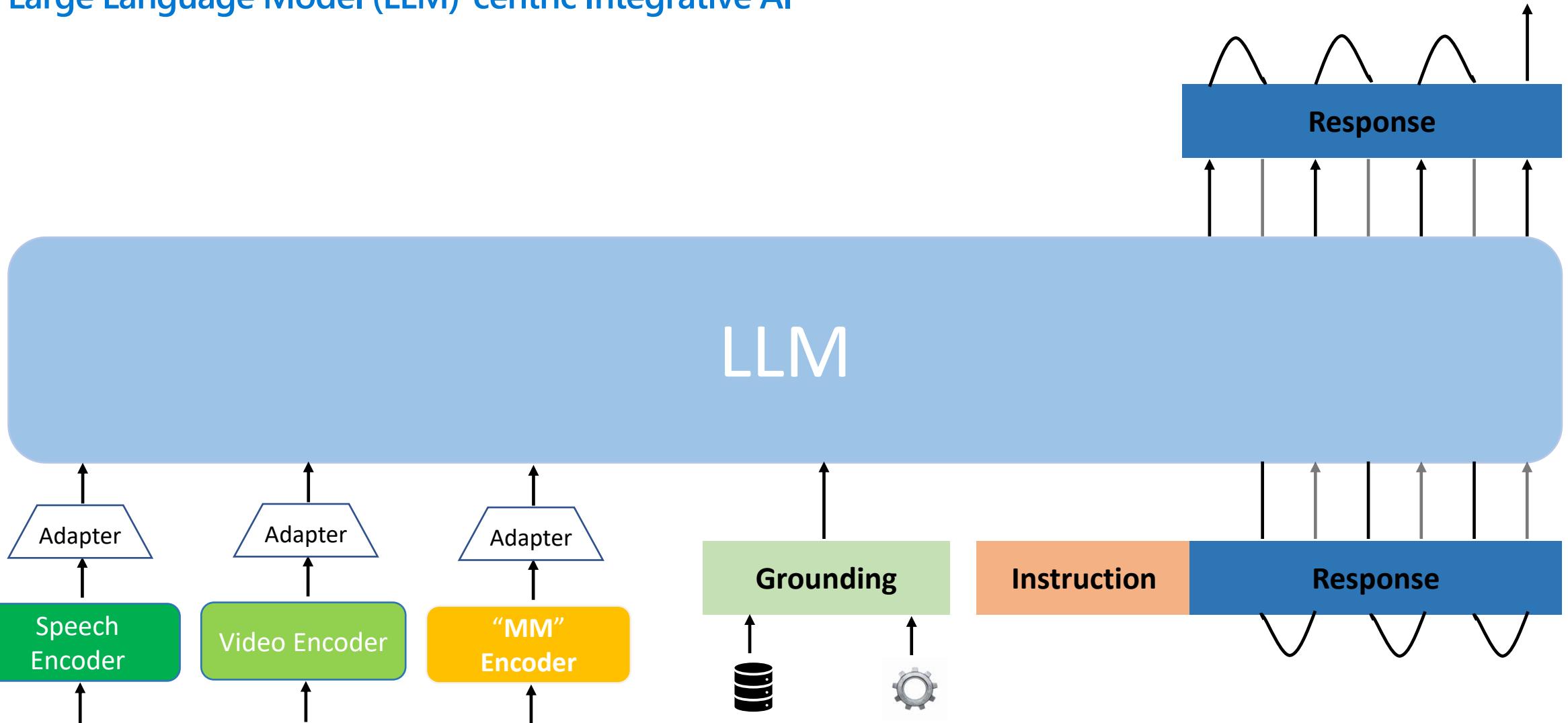


Image credit to XD Huang

What does GPT-V mean to computer vision?



TASK

Logo Recognition



GPT-V

The image shows a close-up of a person's foot wearing a white **Nike Air Force 1 sneaker**. The sneaker is a low-top style with a white rubber sole and white laces. ...

TASK

Object Localization/Detection



GPT-V

There is only one person in the image, so the bounding box for this person would be:
- Top left: (254, 118)
- Bottom right: (731, 693)

TASK

Object Counting



GPT-V

There are **7 apples** in the image.

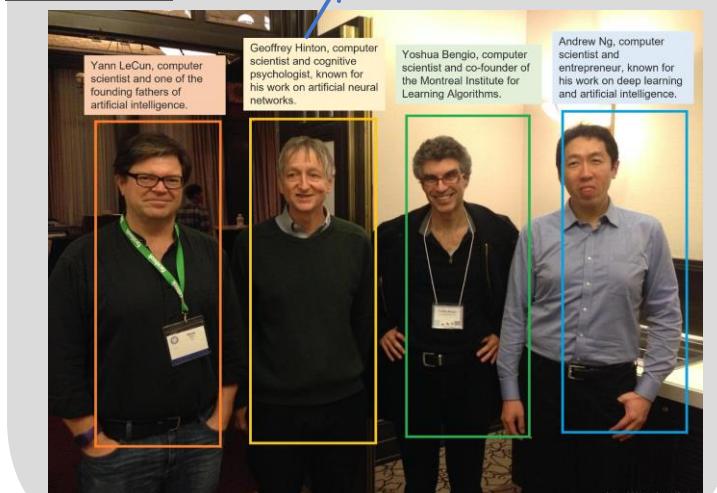
TASK

Dense Captioning



GPT-V

Yang, Z., et al. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv:2309.17421*, 2023.
Image credit to Lijuan Wang



What does GPT-V mean to computer vision?

TASK

Image Recognition



GPT-V

The object in the image is a **traffic light**.

TASK

Celebrity Recognition



GPT-V

The person in the image is **Jensen Huang**, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

TASK

Landmark Recognition



GPT-V

The landmark in the image is the **Space Needle**, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. ...

TASK

Food Recognition



GPT-V

This dish is called a **Cajun Sampler**. It is a sampler platter that consists of four small bowls of traditional Cajun dishes - gumbo, jambalaya, red beans and rice, and crawfish étouffée. ...

What does GPT-V mean to computer vision?

TASK

Scene Text Recognition



GPT-V

The scene text in the image are:

- Vacation - Beach - Smile
- Summer - Watermelon
- Flip flops - Lemonade

TASK

Multilingual Image Description



GPT-V

这是一只小**柯基犬**，它在草地上走着，嘴巴张开，舌头伸出来，好像很开心的样子。它的身上有一条项圈，上面挂着一个**骨头形状的牌子**。

TASK

Document Reasoning

Which country has the highest profit margin?

Country	Sales Volume	Revenue	Profit	Profit Margin
USA	40.080	\$15.971.880	\$3.086.421	19,3%
China	35.070	\$15.866.670	\$3.032.162	19,1%
Australia	27.054	\$14.812.566	\$2.868.636	19,4%
India	23.046	\$10.608.174	\$1.853.710	17,5%
South Korea	16.032	\$10.494.948	\$1.975.844	18,8%
Total / Avg	141.282	\$67.754.238	\$12.816.772	18,8%

GPT-V

Based on the data provided, the country with the highest profit margin is **Australia**, with a profit margin of 19,4%.

TASK

Multilingual Text Recognition



GPT-V

The text in the image is in Chinese and translates to "**Mapo Tofu**." Mapo tofu is a popular Chinese dish made with tofu and minced meat in a spicy sauce.

What does GPT-V mean to computer vision?

TASK

Video Activity Anticipation



GPT-V

The images show the process of making sushi rolls. The first image shows rice being mixed with vinegar in a bowl. The second image shows the rice being spread onto a sheet of seaweed. The third image shows the addition of salmon and cucumber to the roll. The next steps would be to **roll the sushi using a bamboo mat, cut the roll into smaller pieces, and serve with soy sauce and wasabi.**

TASK

Video Action Recognition



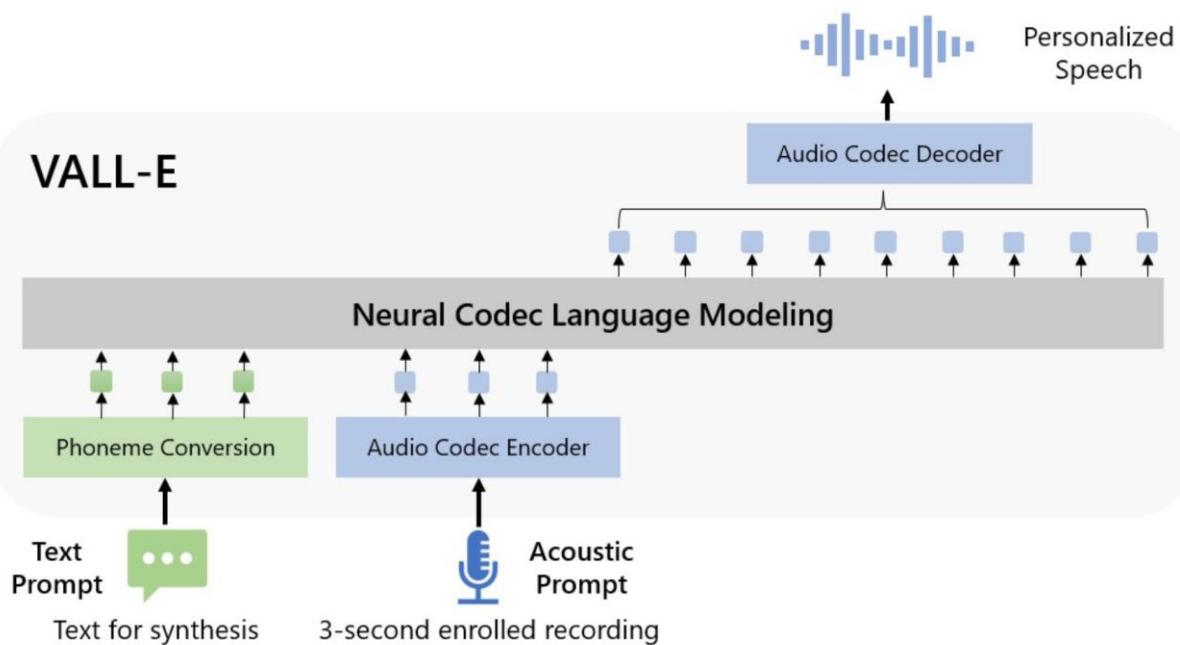
GPT-V

The images show a person in a motion capture suit in a studio with multiple cameras and sensors. The person is **performing a push-up exercise**. In the first image, the person is in the starting position with their arms extended and body in a straight line. In the second image, the person is lowering their body towards the ground. In the third image, the person is at the bottom of the push-up with their chest close to the ground. In the fourth image, the person is pushing their body back up to the starting position.

VALL-E: Neural codec language model

- High quality Zero shot TTS: In context learning through prompts
 - “Steal voice from 3 second's prompt”

Model Overview



Prompt		Output
	I like hamburger but I	
	love noodles much more	

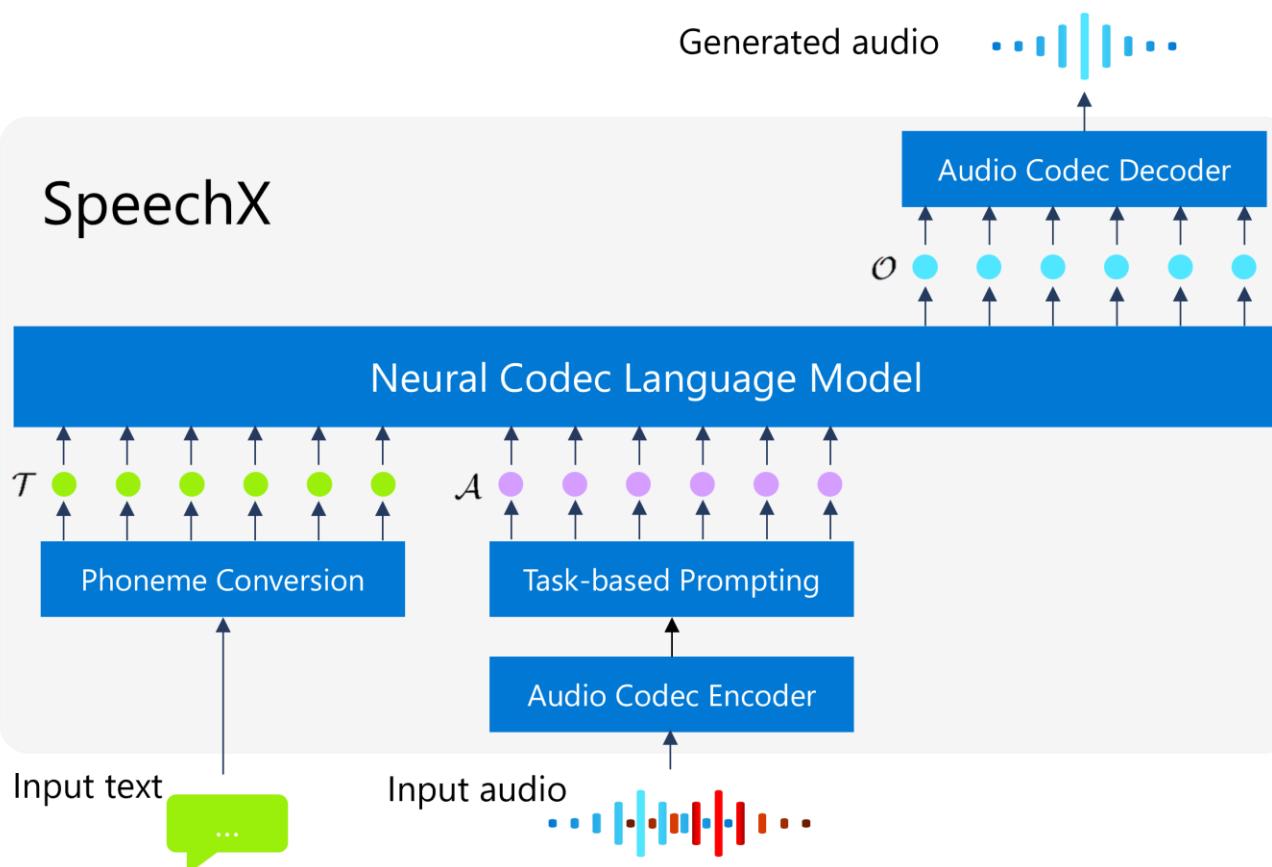
SpeechX – A versatile speech generation model



Versatility: able to handle a wide range of tasks from audio and text inputs.

Robustness: applicable in various acoustic distortions, especially in real-world scenarios where background sounds are prevalent.

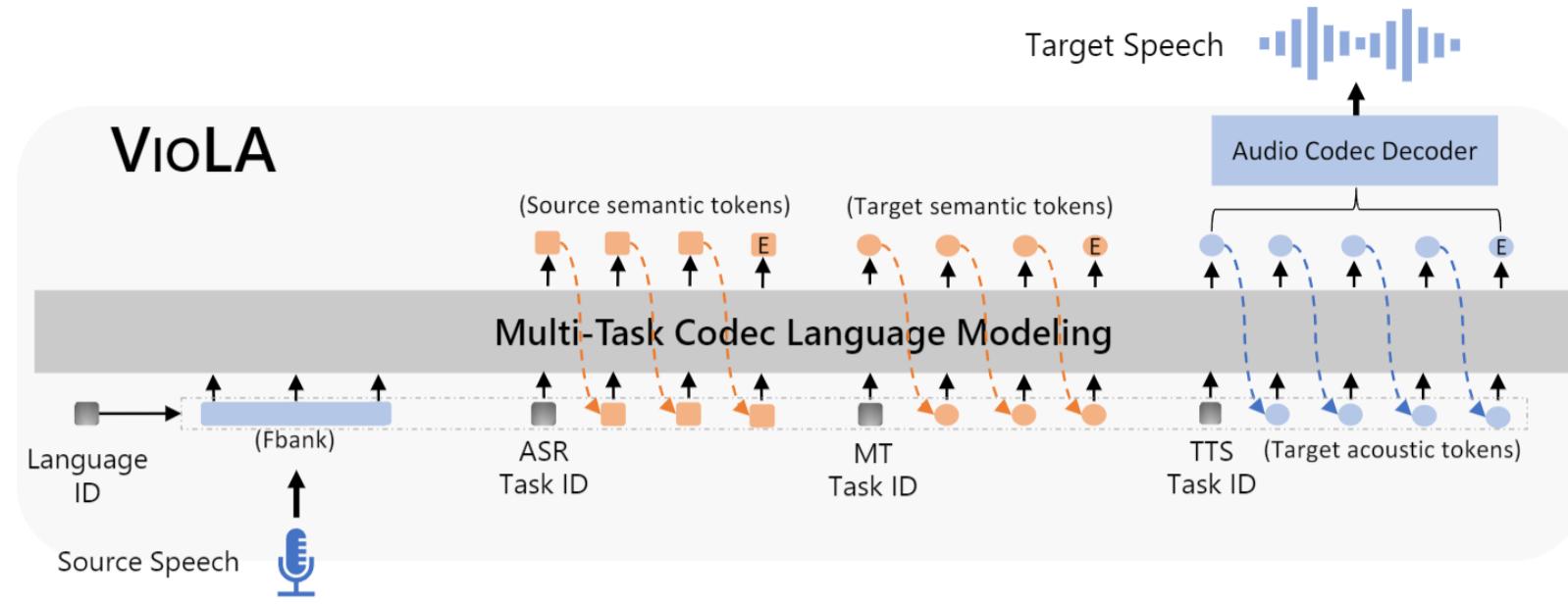
Extensibility: flexible architectures, allowing for seamless extensions of task support.



Task	Input text	Input audio	Output audio
Noise suppression	Transcription (optional)	Noisy speech	Clean speech
Speech removal	Transcription (optional)	Noisy speech	Noise
Target speaker extraction	Transcription (optional)	Speech mixture, Enrollment speech	Clean speech of target speaker
Zero-short TTS	Text for synthesis	Enrollment speech	Synthesized speech mimicking target speaker
Clean speech editing	Edited transcription	Clean speech	Edited speech
Noisy speech editing	Edited transcription	Noisy speech	Edited speech with original background noise

[More demo samples: SpeechX - Microsoft Research](#)

VioLA: A multi-modal model with discrete audio inputs



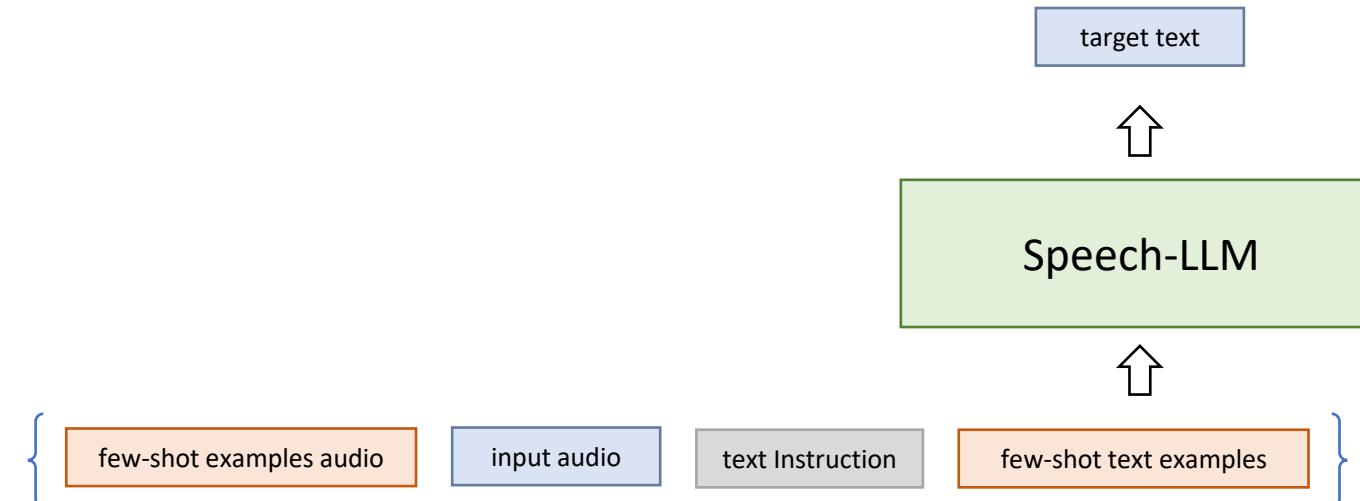
Speech and text can freely serve as input and output

- An extension to audio codec language model
- Naturally merge speech-language tasks
 - Speech recognition
 - Machine translation
 - Speech generation

Input	Output	Typical Tasks
Speech	Text	ASR, ST
Text	Text	MT, LM
Text	Speech	multilingual TTS

Advancing Speech-LLM For In-context Learning

- Trained tasks (EN only)
 - ASR
 - Speech-based Question Answering
- Emergent Capable tasks
 - 0-shot and 1-shot En->X ST
 - 1-shot domain adaptation
 - Instruction-followed ASR



Conclusions

- E2E models are now the mainstreaming ASR models.
 - Streaming Transformer Transducer with masks can achieve very high accuracy and low latency.
- To further advance E2E models, we have discussed several key technologies.
 - Leverage unpaired text: domain adaptation
 - Multilingual: configurable multilingual model
 - Multi-talker ASR: (token-level) serialized output training
 - Speech translation: streaming multilingual speech model
- Large language model (LLM) centric integrative AI may be the next trend.



Thank You!