Recent Advances in End-to-End Automatic Speech Recognition

Jinyu Li

Outline

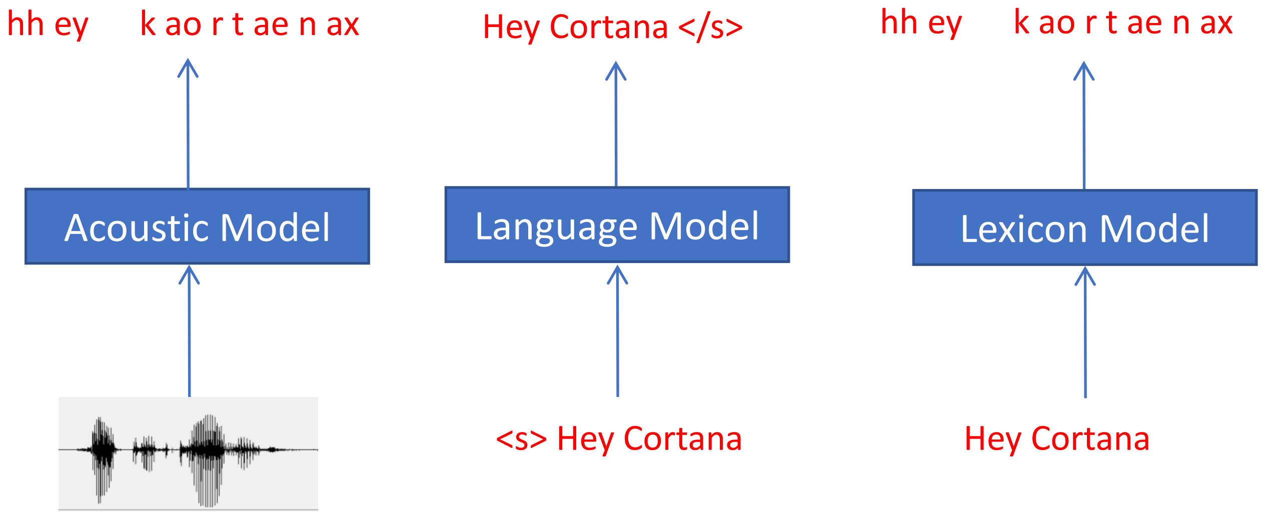
- End-to-end (E2E) automatic speech recognition (ASR) fundamental
- E2E advances
  - Leveraging unpaired text
  - Multilingual ASR
  - Multi-talker ASR
  - Beyond ASR
- The next trend
- Conclusions

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.



hh ey k ao r t ae n ax

Acoustic Model

Hey Cortana </s>

Language Model

<s> Hey Cortana

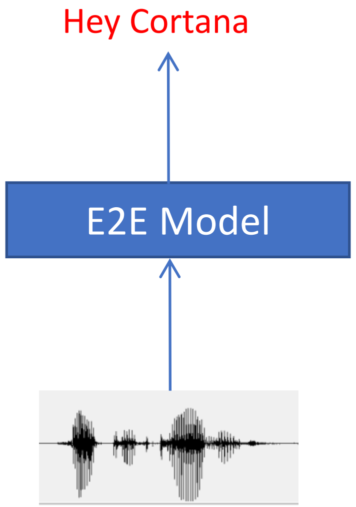
hh ey k ao r t ae n ax

Lexicon Model

Hey Cortana

E2E

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



Hey Cortana

E2E Model

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

Graves and Jaitly, “Towards end-to-end speech recognition with recurrent neural networks” PMLR, 2014.
Hannun et al., “Deep speech: Scaling up end-to-end speech recognition,” in arXiv preprint, 2014.

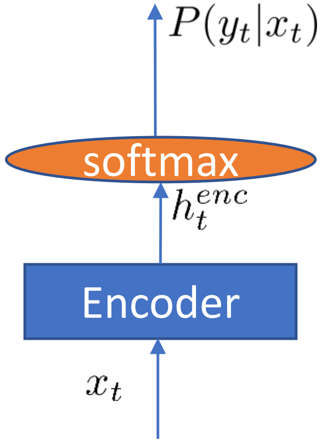
Current Status

- E2E models achieve the state-of-the-art results in most benchmarks in terms of ASR accuracy.

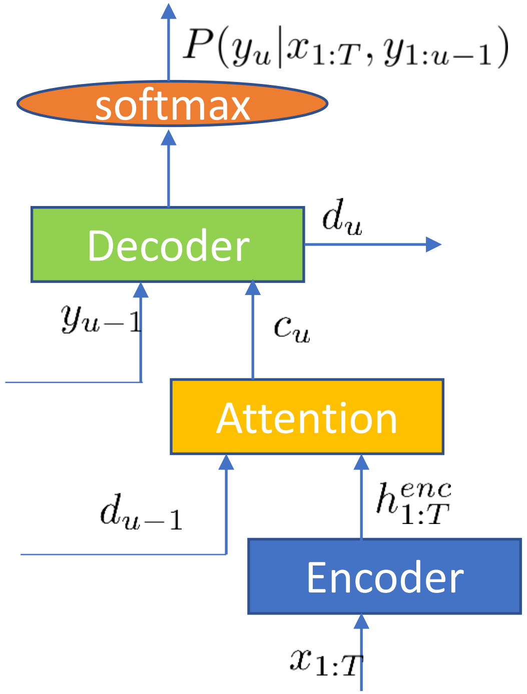
- Practical challenges such as streaming, latency, adaptation capability etc., have been also optimized in E2E models.

- E2E models are now the mainstream models not only in academic but also in industry.

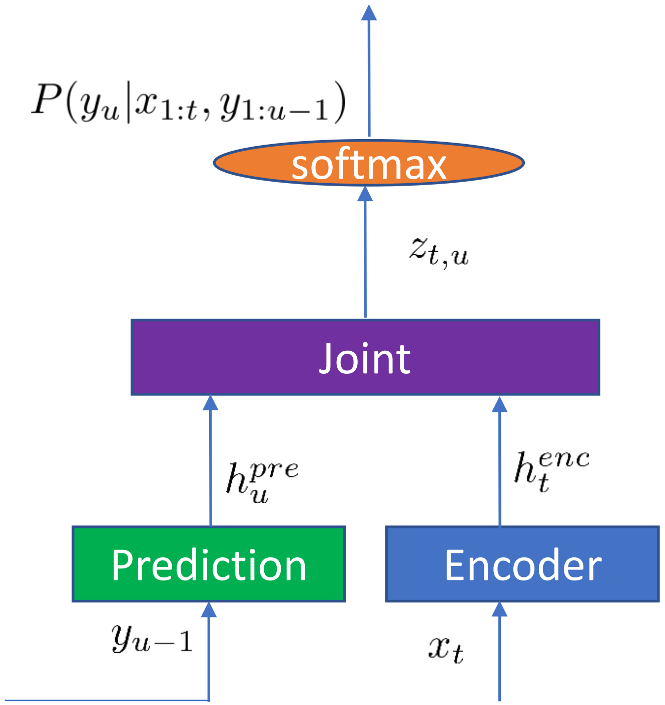
E2E Models



Connectionist Temporal Classification (CTC)



Attention-based encoder decoder (AED)



RNN-Transducer (RNN-T)

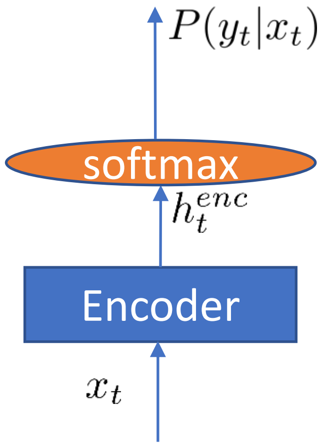
E2E Models

	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

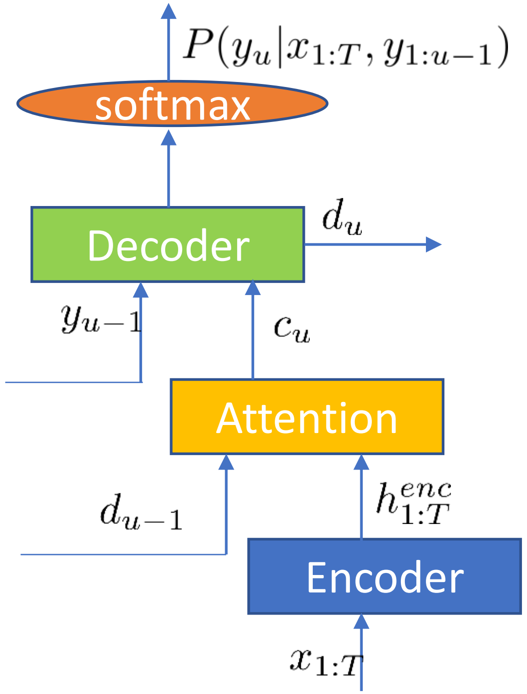
RNN-T is the most popular E2E model in industry which requires streaming ASR.

Sainath, T., et al. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. in *Proc. ICASSP*, 2020
Li, J., et al., Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability. in *Proc. Interspeech,* 2020.

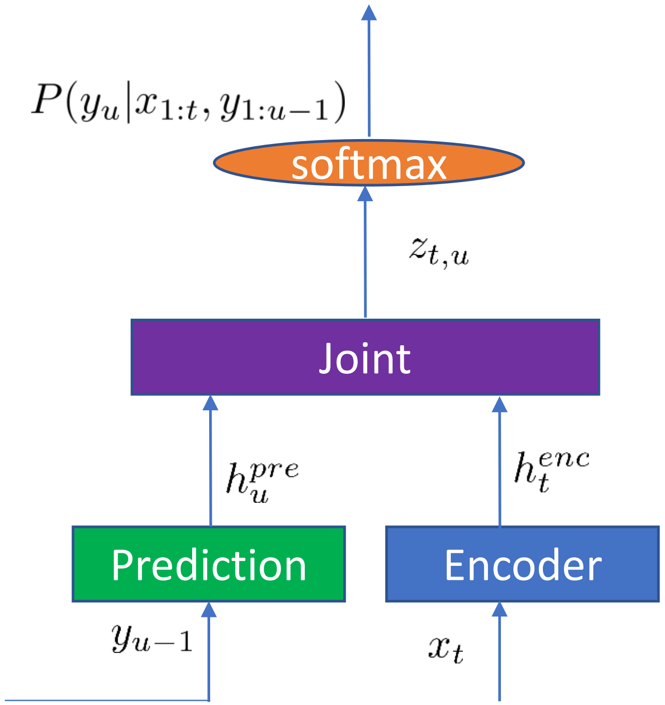
Encoder is the Most Important Component



Connectionist Temporal Classification (CTC)

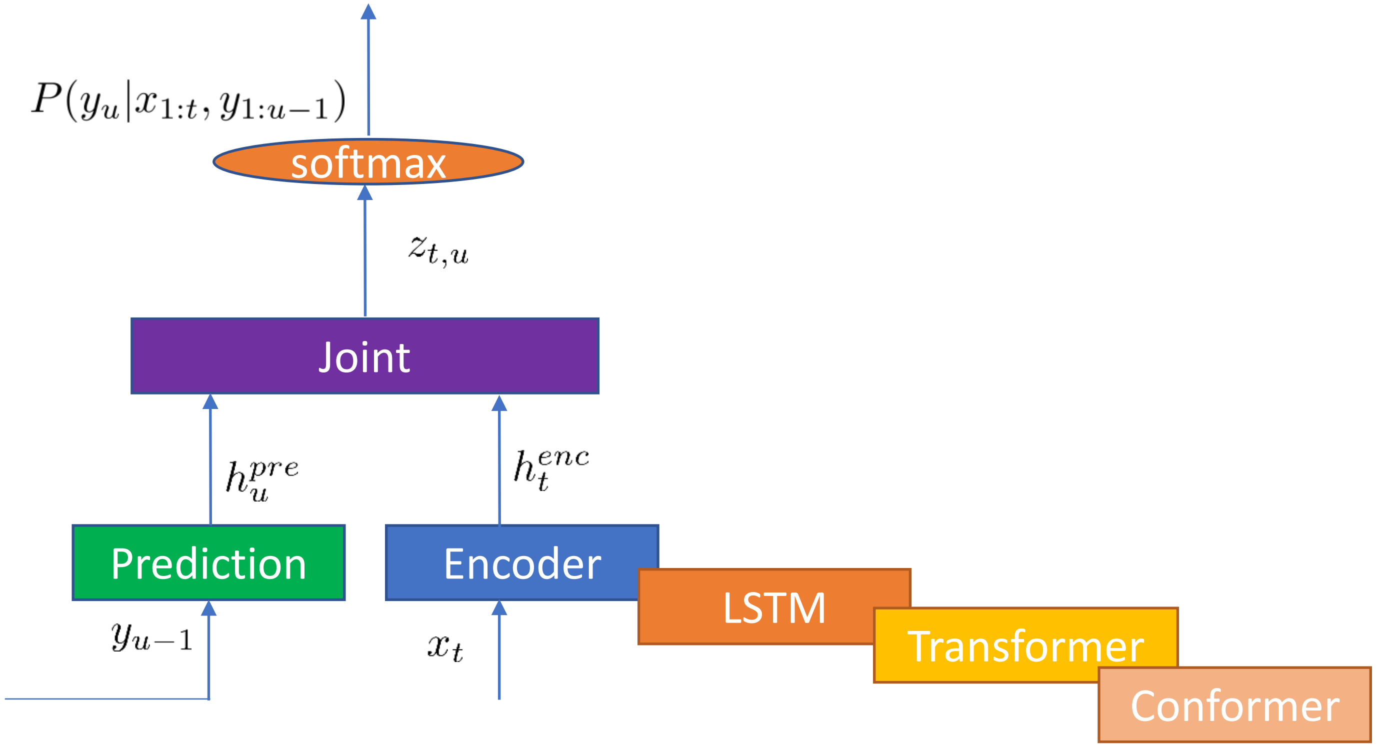


Attention-based encoder decoder (AED)



RNN-Transducer (RNN-T)

Encoder for RNN-T



Transformer

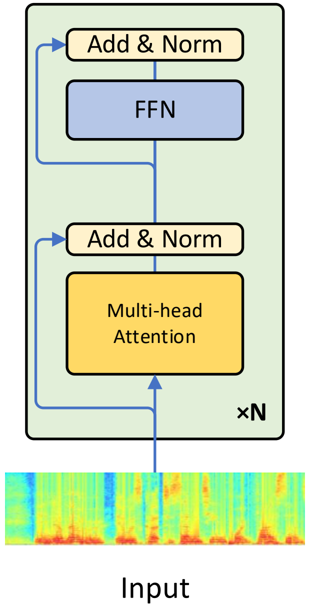
- Self-attention: computes the attention distribution over the input speech sequence

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

- 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



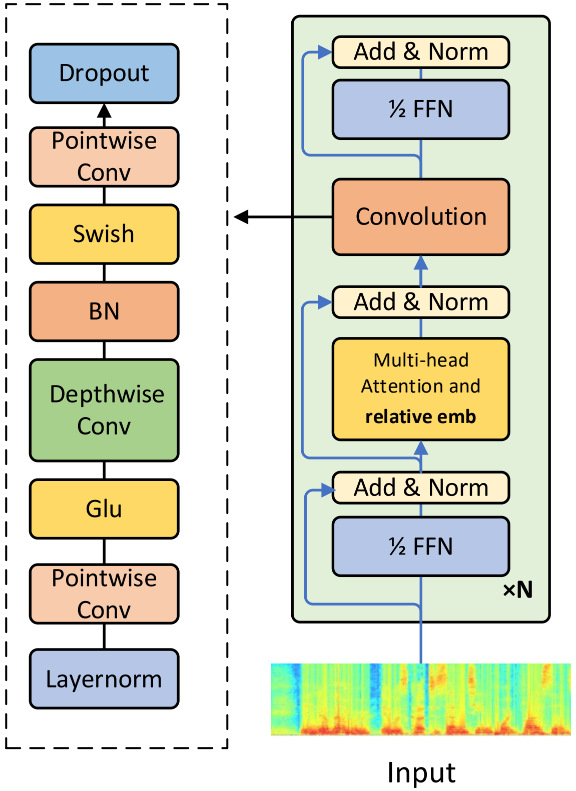
Vaswani et al. "Attention is all you need" NIPS 2017
Zhang et al., "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," in Proc. ICASSP 2020.

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 $\{\boldsymbol{x}_\tau\}$, binary attention mask $\{m_{t,\tau}\}$ to control range of input $\{\mathbf{x}_\tau\}$ to use

$$\alpha_{t,\tau} = \frac{\mathrm{m}_{t,\tau}\exp(\beta\left(W_q\boldsymbol{x}_t\right)^T(W_k\boldsymbol{x}_\tau))}{\sum_{\tau'}\mathrm{m}_{t,\tau'}\exp(\beta\left(W_q\boldsymbol{x}_t\right)^T(W_k x_{\tau'}))} = softmax(\beta\boldsymbol{q}_t^T\boldsymbol{k}_\tau, m_{t,\tau})$$

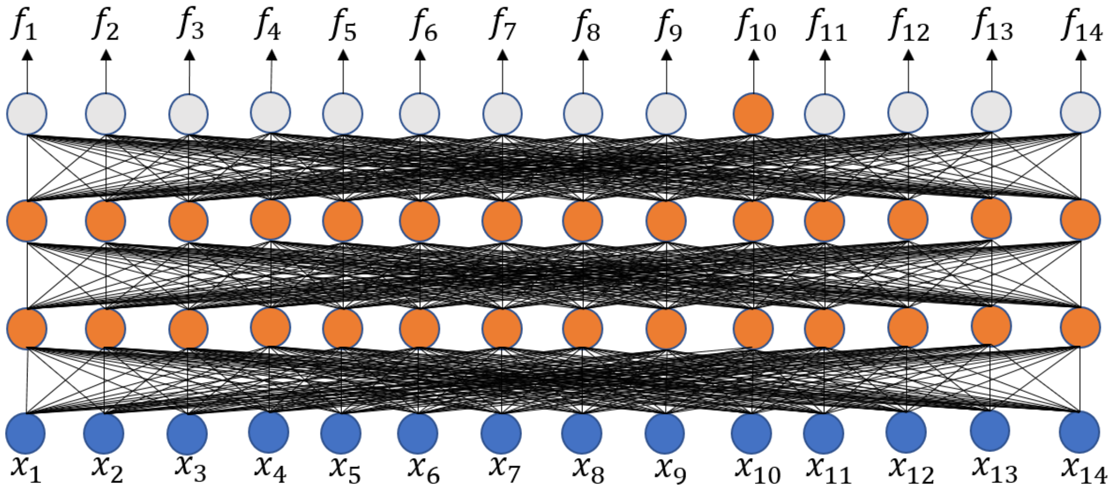
- Apply attention weight over value vector $\{\boldsymbol{v}_\tau\}$

$$z_t = \sum_\tau \alpha_{t,\tau}W_v\boldsymbol{x}_\tau = \sum_\tau \alpha_{t,\tau}\boldsymbol{v}_\tau$$

Chen, X., et al. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. in *Proc. ICASSP,* 2021.

Attention Mask is All You Need

- Offline (whole utterance)



$$\begin{array}{cc} \begin{array}{c}\text{Frame}\\ \text{Index}\end{array} & \\ \begin{array}{c} 1\\2\\3\\4\\5\\6\\7\\8\\9\\10\\11\\12\\13\\14 \end{array} & \begin{bmatrix} 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1 \end{bmatrix}. \end{array}$$

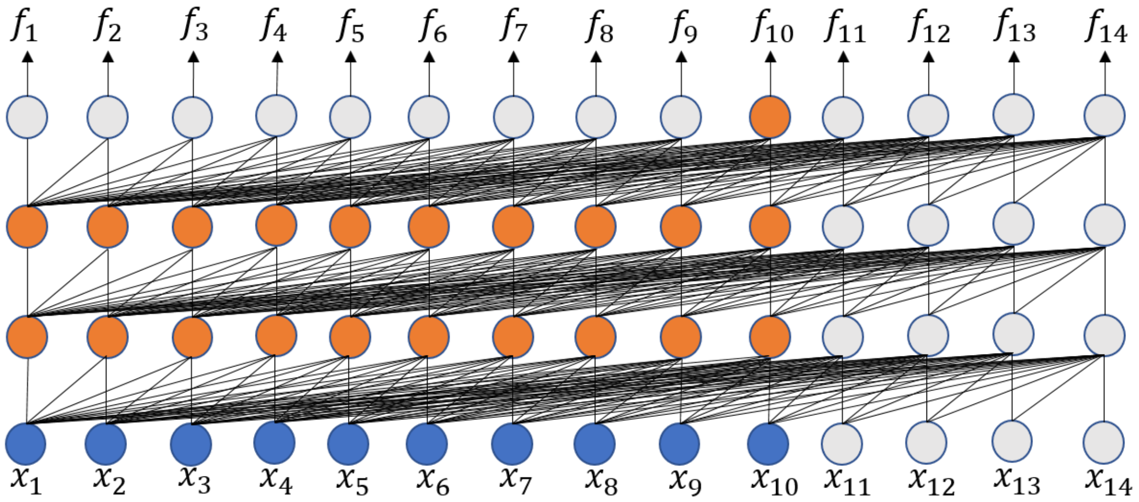
Predicting output for x_{10}

Not streamable

Attention Mask

Attention Mask is All You Need

- 0 lookahead, full history



Predicting output for x_{10}

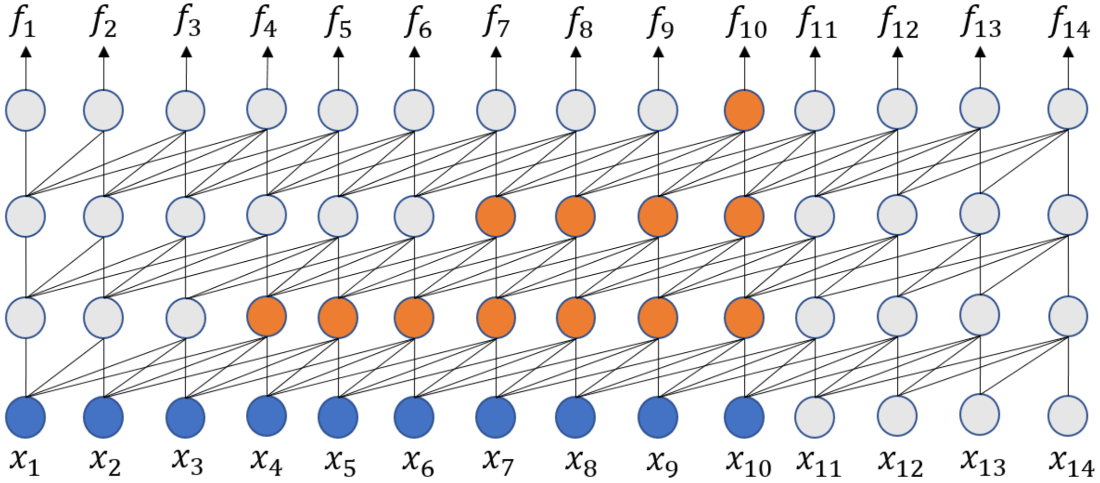
$$\begin{array}{c|cccccccccccccc}
\text{Frame Index} & & & & & & & & & & & & & & \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
9 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
10 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
11 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
12 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
13 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
14 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array} .$$

Attention Mask

Memory and runtime cost increase linearly

Attention Mask is All You Need

- 0 lookahead, limited history (3 frames)


f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10} f_{11} f_{12} f_{13} f_{14}
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13} x_{14}

Predicting output for x_{10}

Frame Index

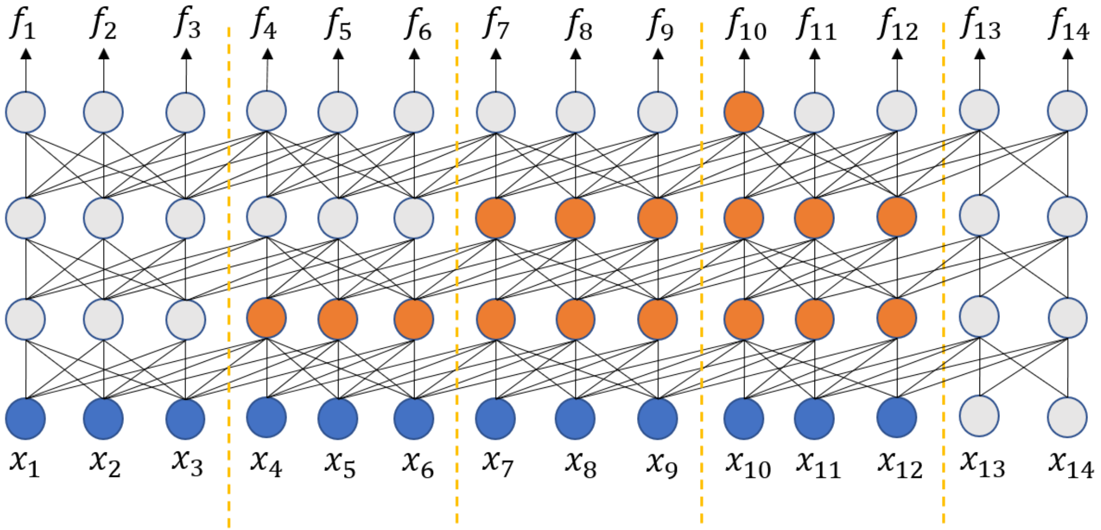
$$
\begin{array}{c}
1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14
\end{array}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{bmatrix}.
$$

Attention Mask

In some scenario, small amount of latency is allowed

Attention Mask is All You Need

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



f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10} f_{11} f_{12} f_{13} f_{14}
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13} x_{14}

Predicting output for x_{10}

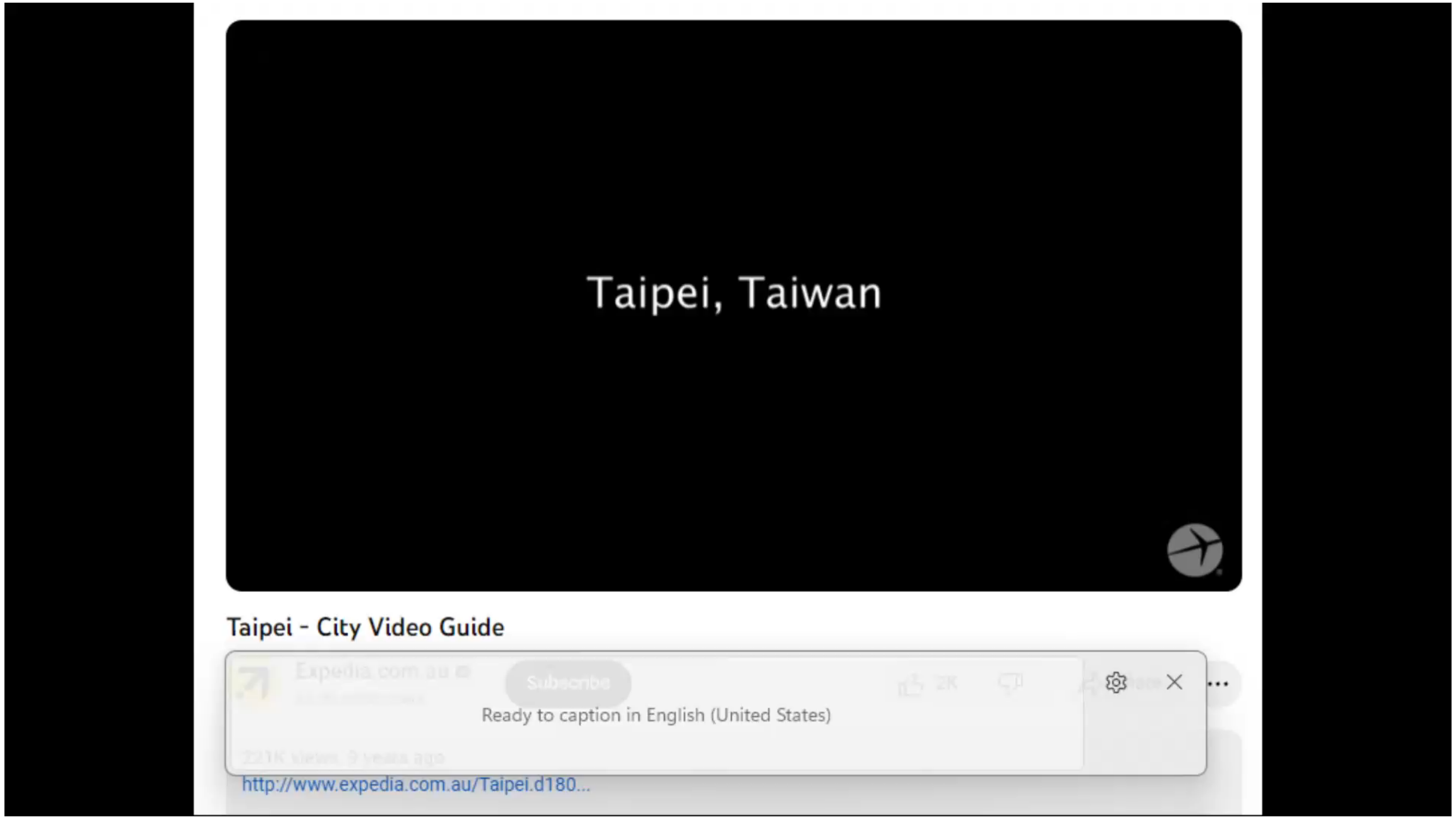
Look-ahead window [0, 2]

Frame Index

$$
\begin{array}{c}
1\\2\\3\\4\\5\\6\\7\\8\\9\\10\\11\\12\\13\\14
\end{array}
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}.
$$

Attention Mask

Live Caption in Windows 11



Taipei, Taiwan

Taipei - City Video Guide

Ready to caption in English (United States)

http://www.expedia.com.au/Taipei.d180...

Advancing E2E Models

- unpaired text
- multi-talker
- multilingual
- speech translation

Unpaired Text

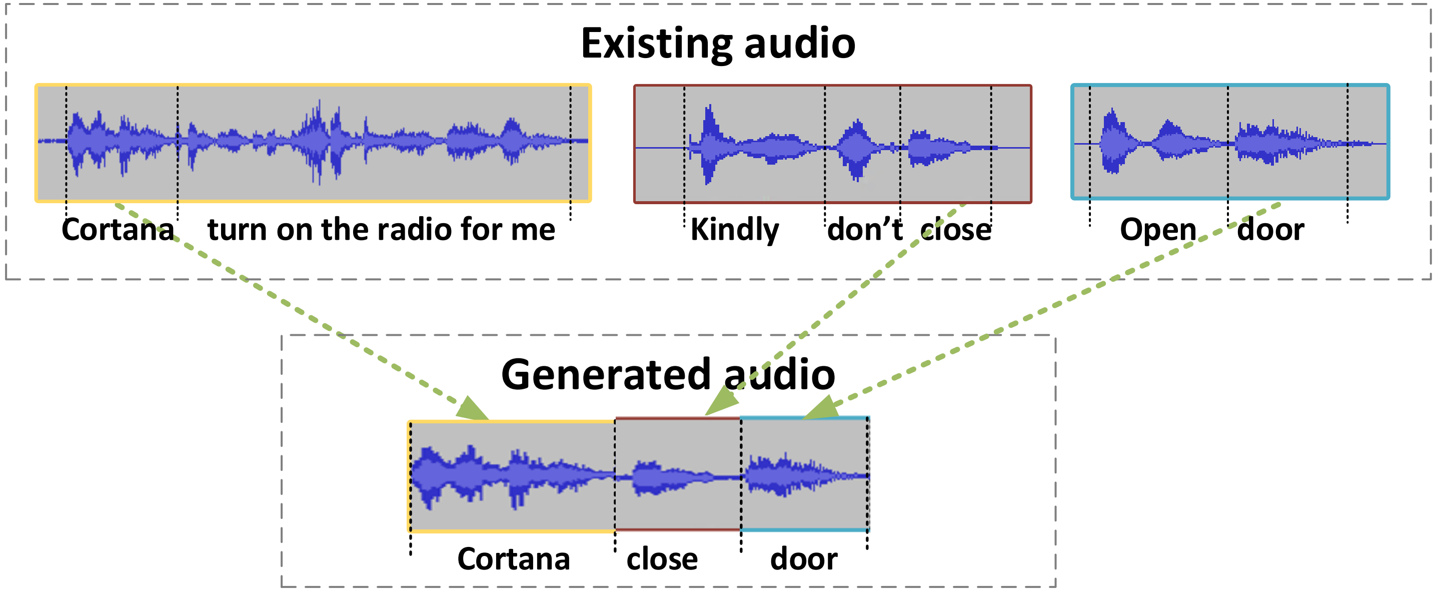
Leverage Unpaired Text

- 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.

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}}\Big[\ \underbrace{\log P\big(\boldsymbol{Y}|\boldsymbol{X};\theta_{\mathrm{E2E}}^{\mathrm{S}}\big)}_{\text{E2E score}} + \lambda_T\ \underbrace{\log P\big(\boldsymbol{Y};\theta_{\mathrm{LM}}^{\mathrm{T}}\big)}_{\text{Target LM score}}\ \Big]$$

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

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}}\Big[\ \underbrace{\log P\big(\boldsymbol{Y}|\boldsymbol{X};\theta_{\mathrm{E2E}}^{\mathrm{S}}\big) + \lambda_T\ \log P\big(\boldsymbol{Y};\theta_{\mathrm{LM}}^{\mathrm{T}}\big)}_{\text{Shallow Fusion score}}\ -\ \underbrace{\lambda_S\ \log P\big(\boldsymbol{Y};\theta_{\mathrm{LM}}^{\mathrm{S}}\big)}_{\text{Source LM score}}\Big]$$

A standalone LM trained with training transcript of E2E model

- HAT/ILME-based Fusion
  - ➢ **Subtract internal LM score** from Shallow Fusion score.

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}}\Big[\underbrace{\log P\big(\boldsymbol{Y}|\boldsymbol{X};\theta_{\mathrm{E2E}}^{\mathrm{S}}\big) + \lambda_T\ \log P\big(\boldsymbol{Y};\theta_{\mathrm{LM}}^{\mathrm{T}}\big)}_{\text{Shallow Fusion score}}\ -\ \lambda_I \underbrace{\log P\big(\boldsymbol{Y};\theta_{\mathrm{E2E}}^{\mathrm{S}}\big)}_{\text{Internal LM score}}\Big]$$

An inherent LM estimated by E2E model parameters

  - ➢ Show **improved** ASR performance over Shallow Fusion and Density Ratio

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.

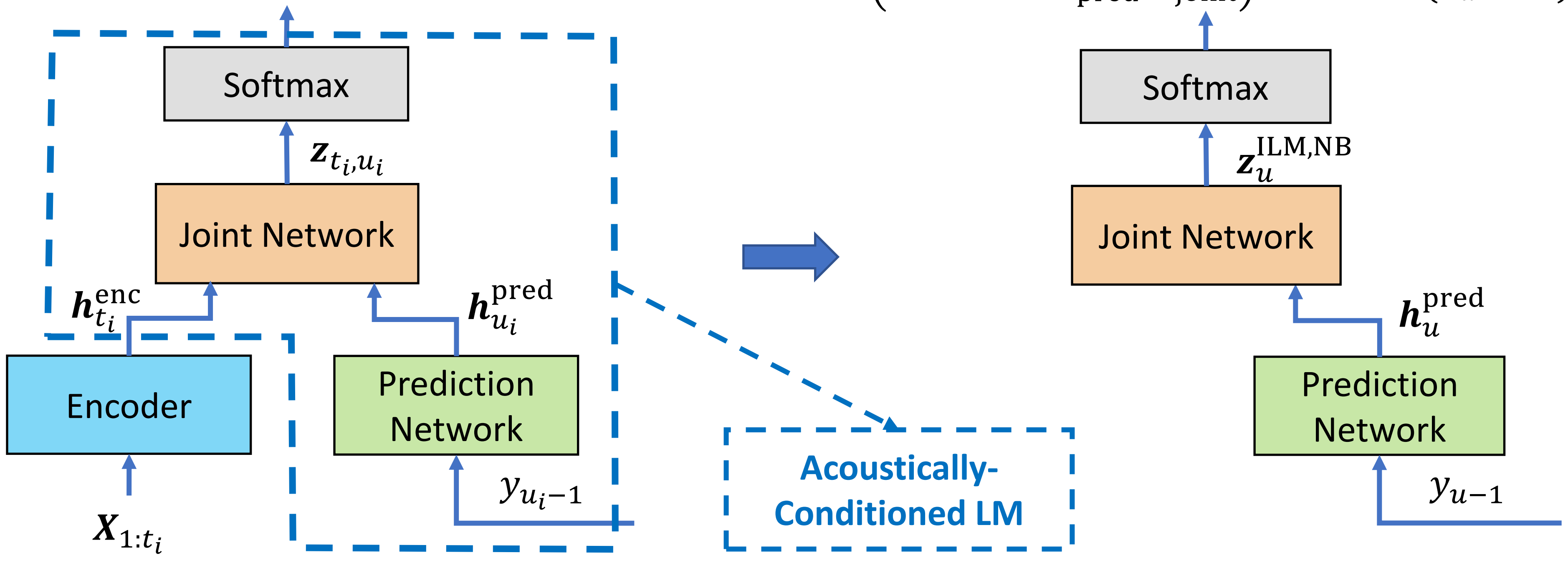
Internal LM Estimation

- RNN-T

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

- **Internal LM estimation of RNN-T**

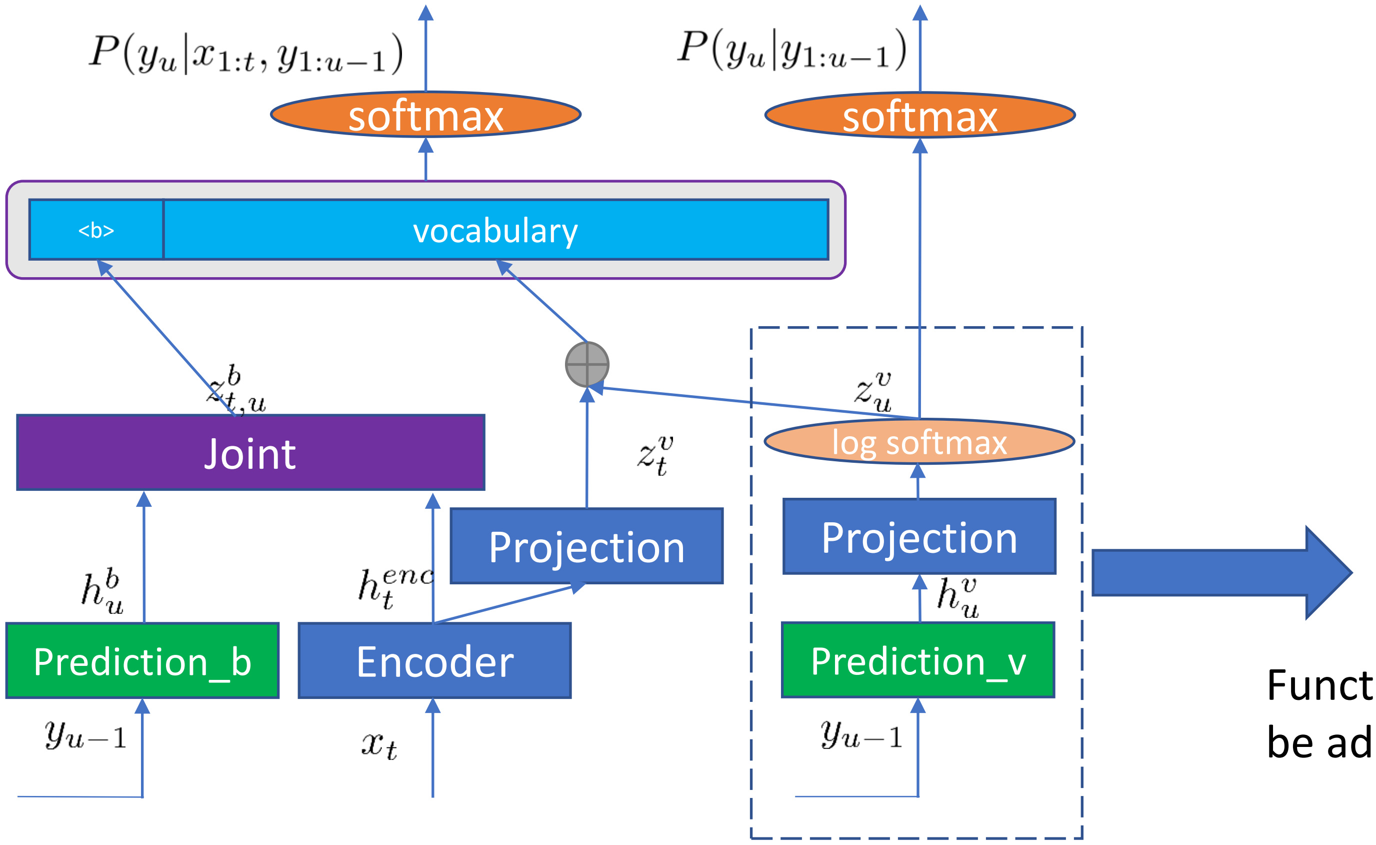
$$P\left(y_u|\boldsymbol{Y}_{0:u-1};\theta_{\text{pred}},\theta_{\text{joint}}\right) = \text{softmax}(\boldsymbol{z}_u^{\text{ILM,NB}})$$


Softmax
$\boldsymbol{z}_{t_i,u_i}$
Joint Network
$\boldsymbol{h}_{t_i}^{\text{enc}}$
$\boldsymbol{h}_{u_i}^{\text{pred}}$
Encoder
Prediction Network
$\boldsymbol{X}_{1:t_i}$
y_{u_i-1}
Acoustically-Conditioned LM
Softmax
$\boldsymbol{z}_u^{\text{ILM,NB}}$
Joint Network
$\boldsymbol{h}_u^{\text{pred}}$
Prediction Network
y_{u-1}

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

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

Factorized Neural Transducer



$$\mathcal{J}_f = \mathcal{J}_t - \lambda \log P(\mathbf{y}_1^U)$$

Functions as a neural LM. Can be adapted with text only data!

Chen, X., et al., Factorized Neural Transducer for Efficient Language Model Adaptation. in *Proc. ICASSP*, 2022.

Multilingual ASR

Multilingual

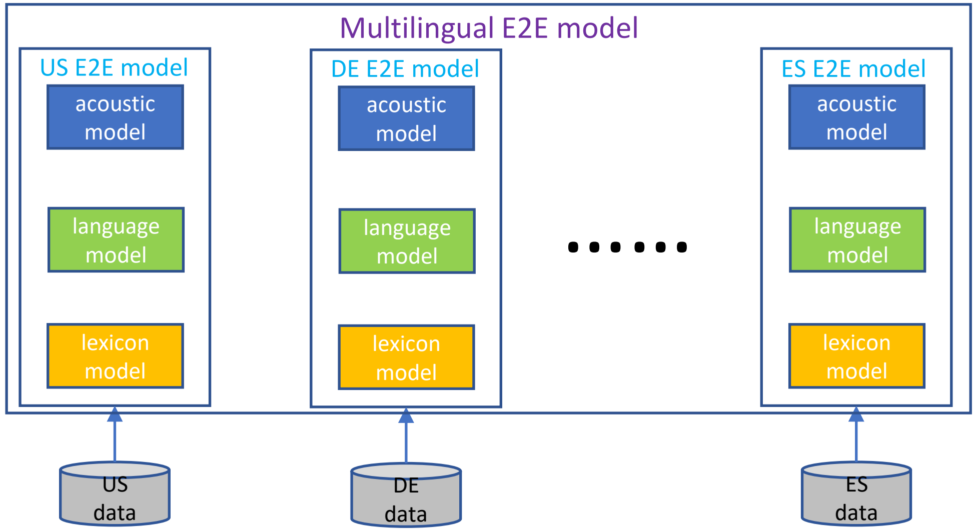
- 40% people can speak only 1 language fluently.
- 43% people can speak only 2 languages fluently.
- 13% people can speak only 3 languages fluently.
- 3% people can speak only 4 languages fluently.
- <0.1% people can speak 5+ languages fluently.

- Human cannot recognize all languages. Can we build a *single high quality multilingual model* to serve *all users*?

Statistics are from: http://ilanguages.org/bilingual.php

Multilingual E2E Models

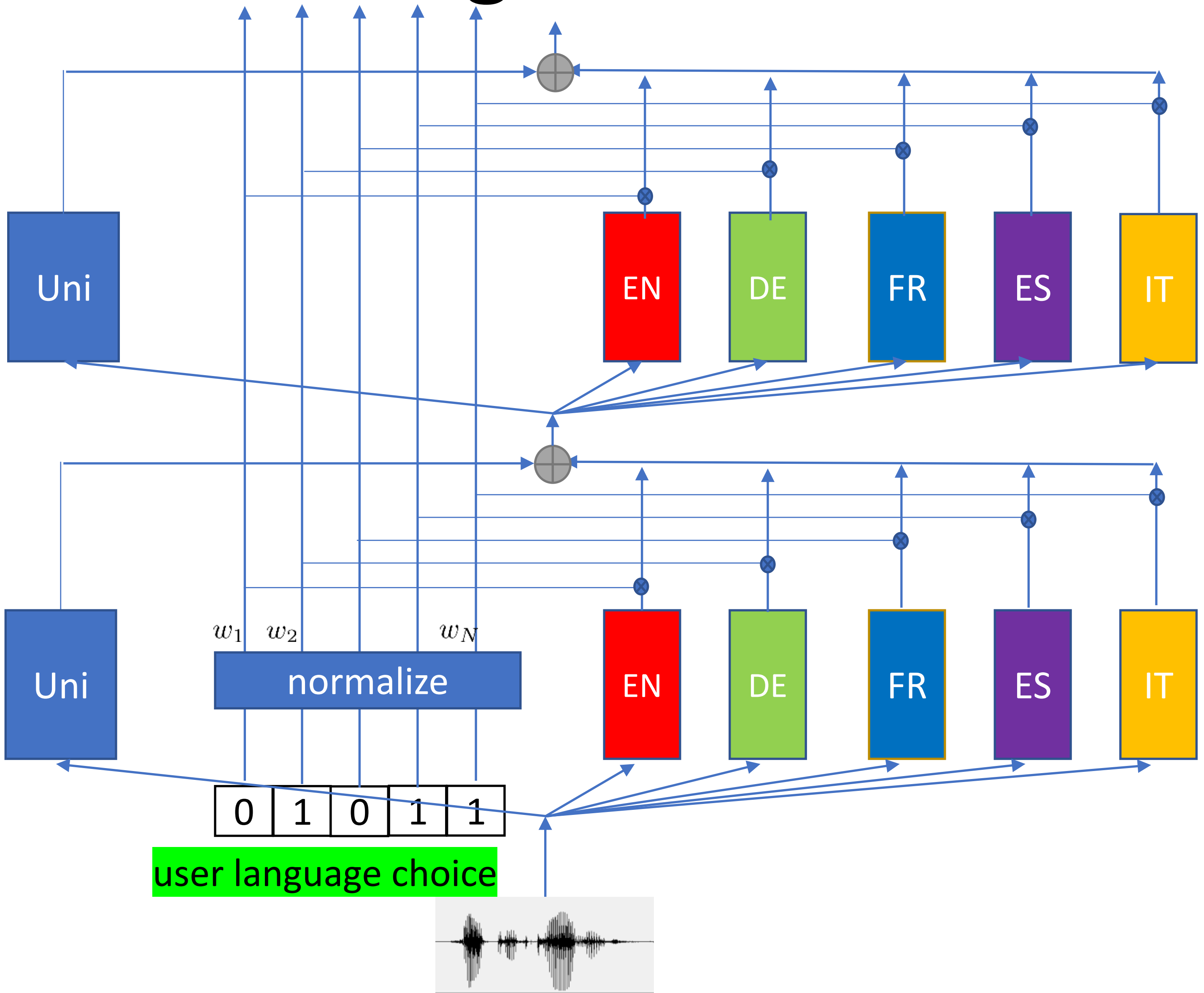
- Double-edged sword of pooling all language data
  - Maximum sharing between languages; One model for all languages
  - Confusion between languages



Configurable Multilingual Model

- **Universal module**: modeling the sharing across languages

- **Expert module**: modeling the residual from universal module for each language



Multi-talker ASR

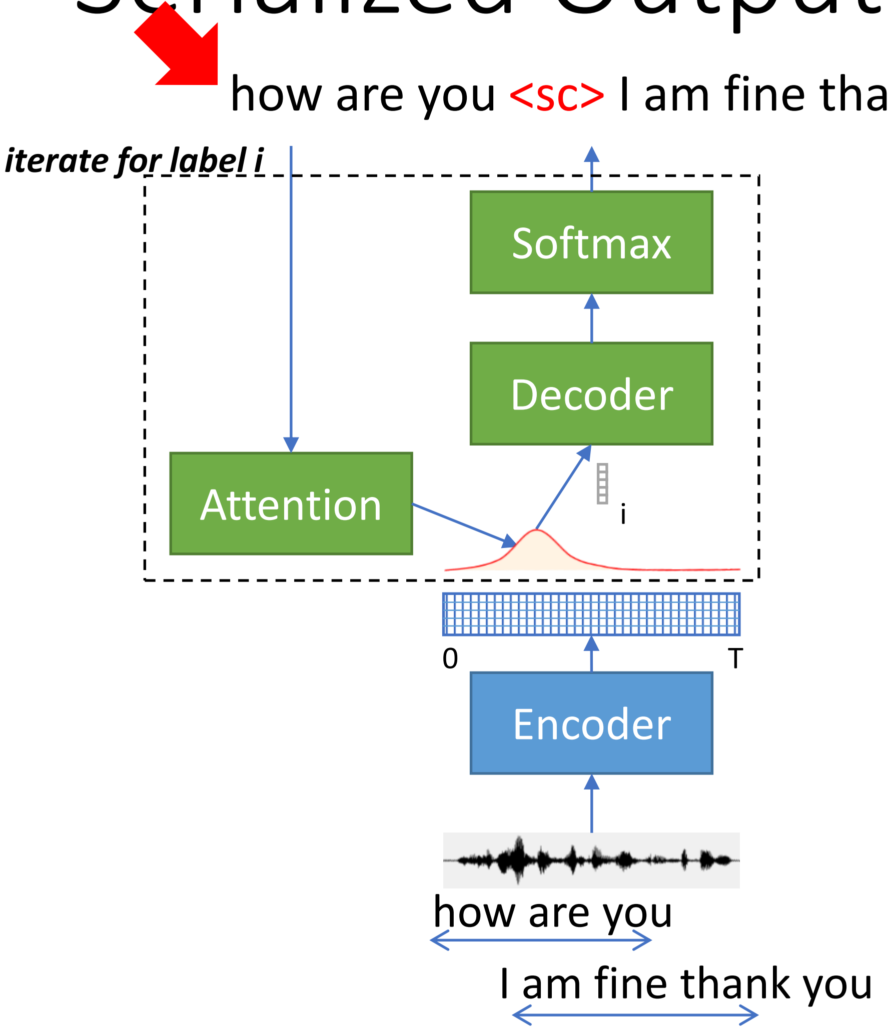
Multi-talker ASR

- 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 ☹

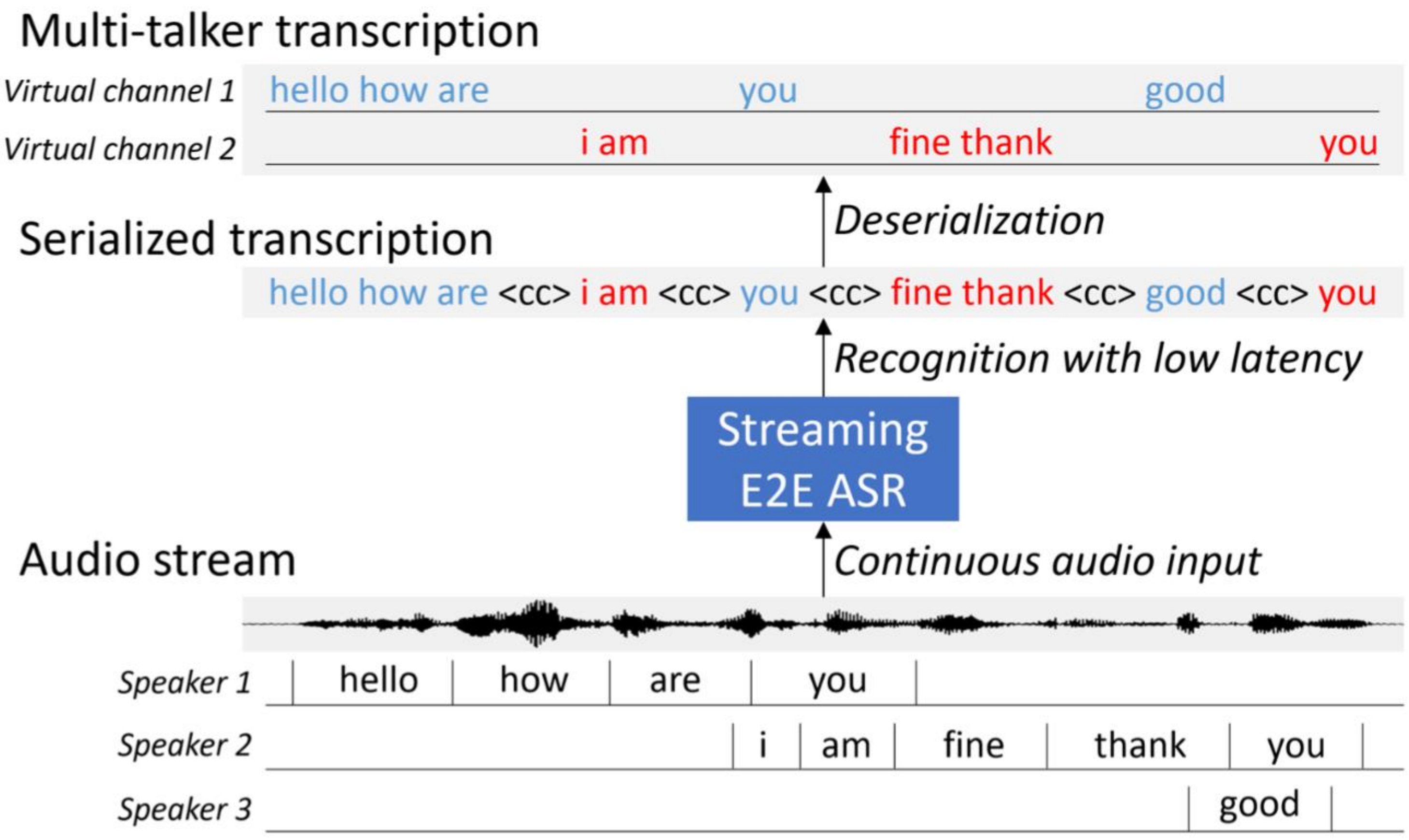
- Solutions: E2E multi-talker models

Serialized Output Training (SOT)



Kanda, N., et al. Serialized Output Training for End-to-End Overlapped Speech Recognition. In Proc. Interspeech, 2020.

Token-level Serialized Output Training (t-SOT)

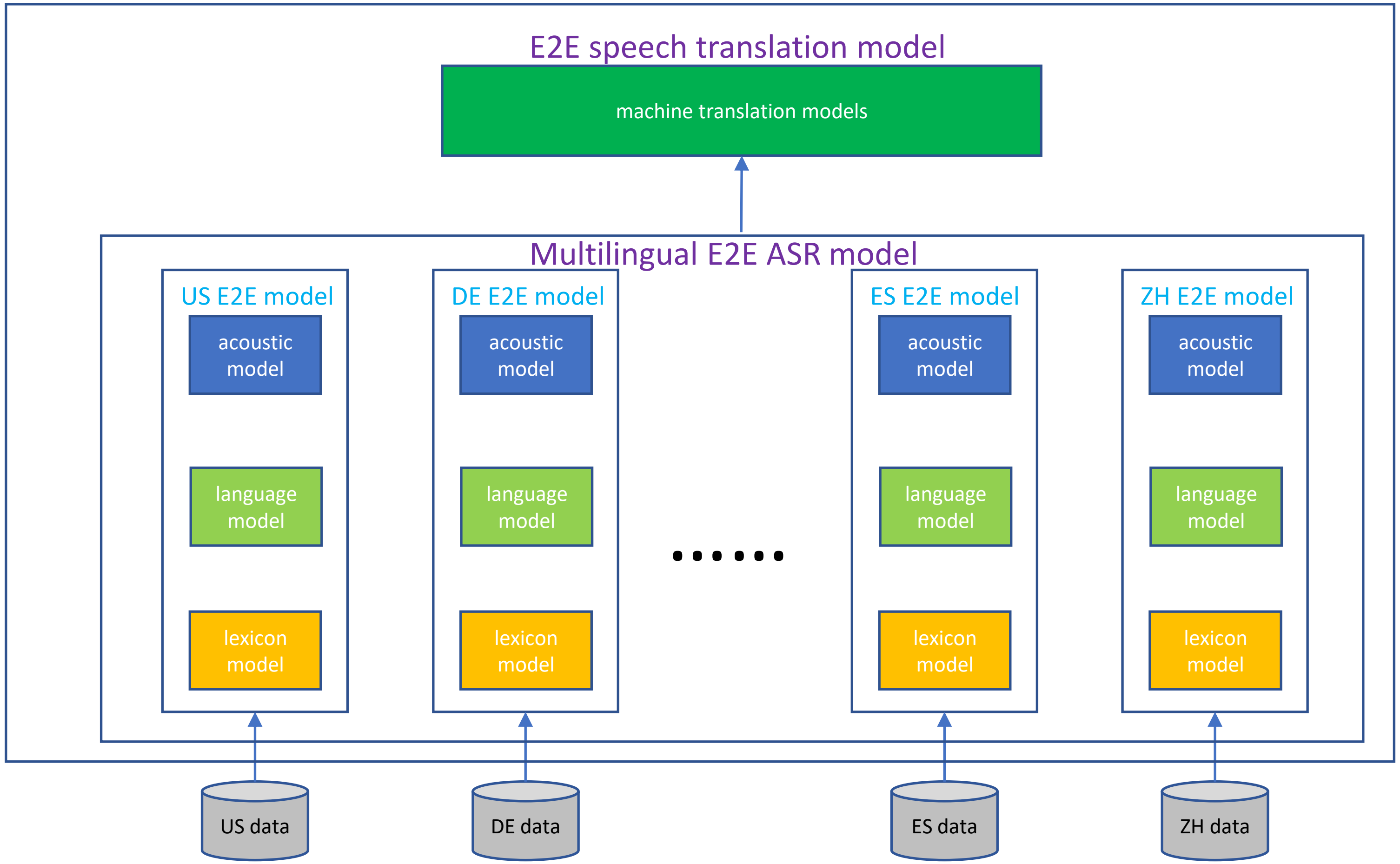


Kanda, N., et al. Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings. in Proc. Interspeech, 2022.

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 **M**ultilingual **S**peech **M**odel (SM^2)

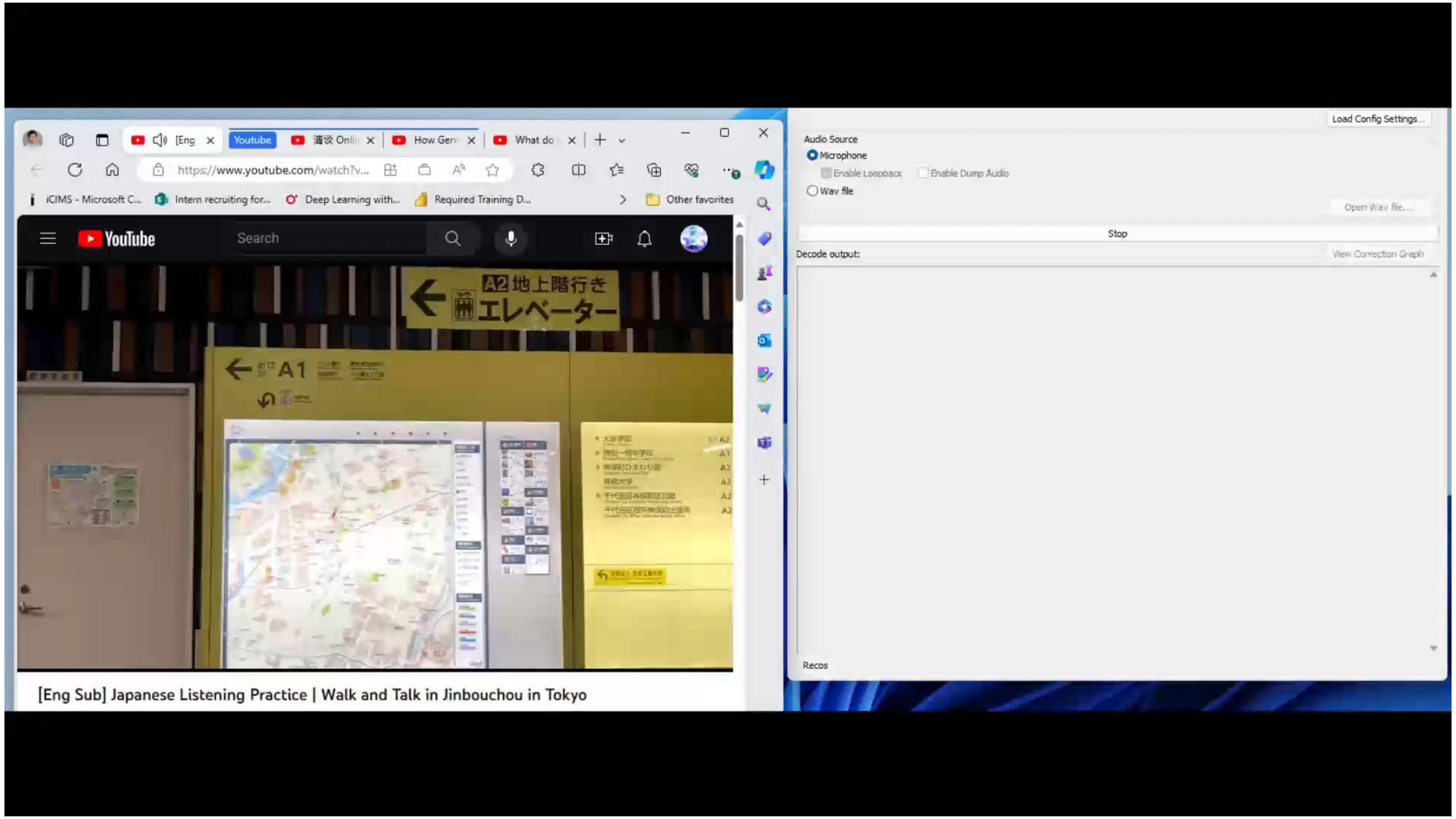
- Multilingual data is pooled together to train a streaming model to perform both ST and ASR functions.

- ST training is totally weakly supervised without using any human labeled parallel corpus.

- The model is very small, running on devices.

Xue, J., et al. A Weakly-Supervised Streaming Multilingual Speech Model with Truly Zero-Shot Capability. In *Proc. ASRU,* 2023.

Simultaneous ST Demo



[Eng Sub] Japanese Listening Practice | Walk and Talk in Jinbouchou in Tokyo

Audio Source
● Microphone
□ Enable Loopback □ Enable Dump Audio
○ Wav file

Decode output:

Recos

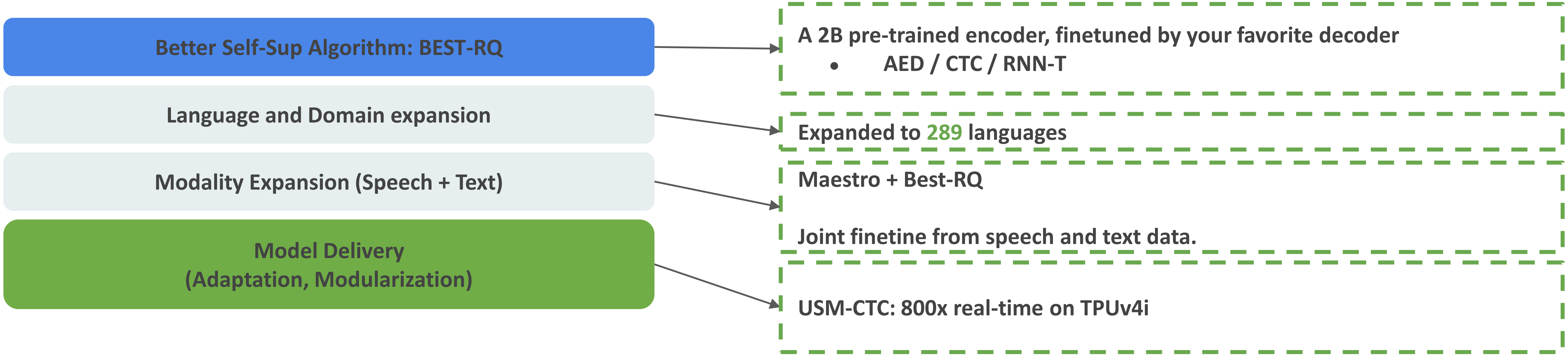
Foundation Models

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

Radford, A., et al. Robust speech recognition via large-scale weak supervision, arXiv:2212.04356, 2022.

Universal Speech Understanding (USM) model



Better Self-Sup Algorithm: BEST-RQ

Language and Domain expansion

Modality Expansion (Speech + Text)

Model Delivery
(Adaptation, Modularization)

A 2B pre-trained encoder, finetuned by your favorite decoder

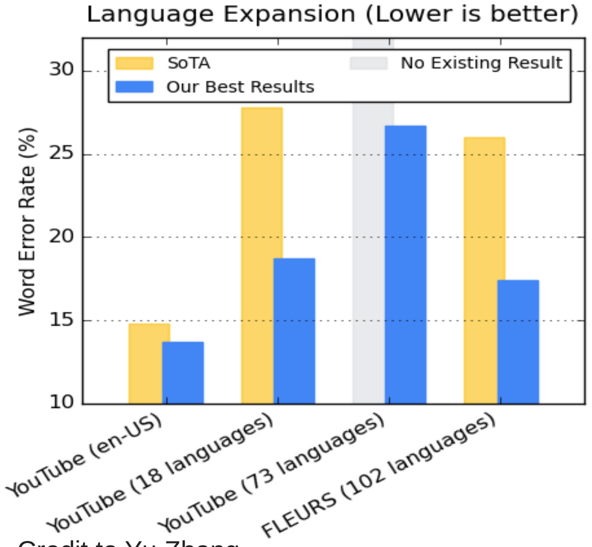
- AED / CTC / RNN-T

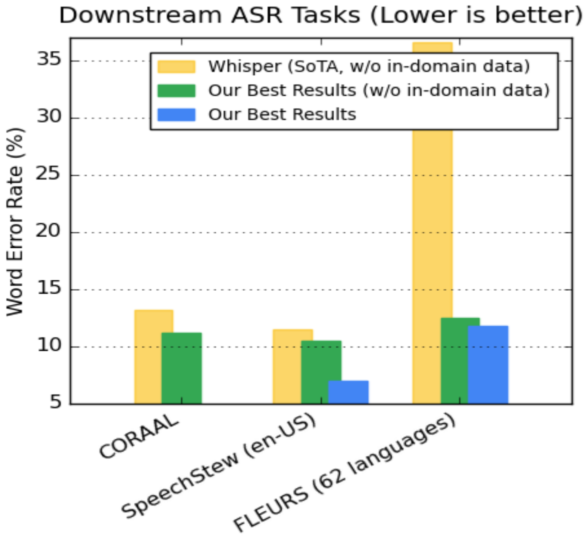
Expanded to 289 languages

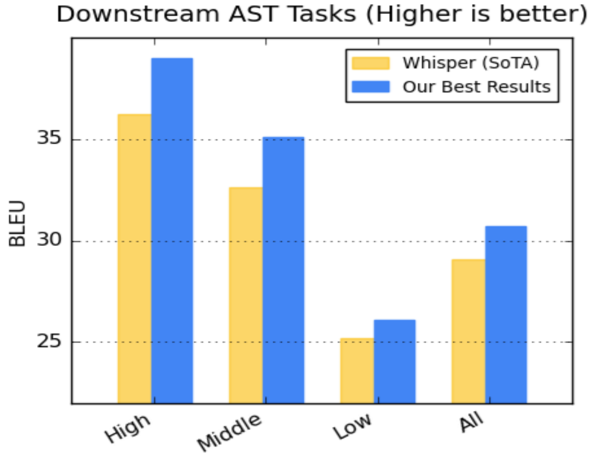
Maestro + Best-RQ

Joint finetine from speech and text data.

USM-CTC: 800x real-time on TPUv4i







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

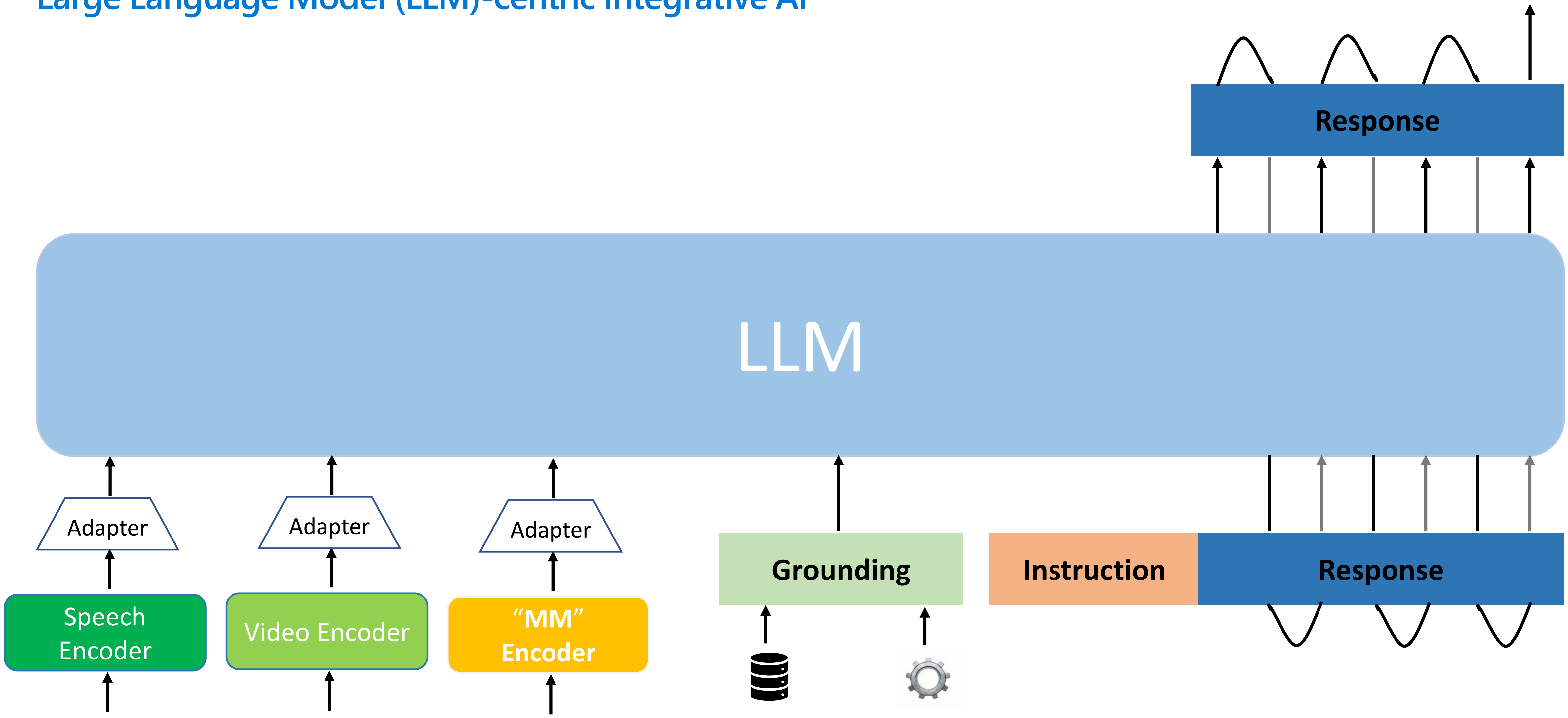


Image credit to XD Huang

What does GPT-V mean to computer vision?

TASK Logo Recognition

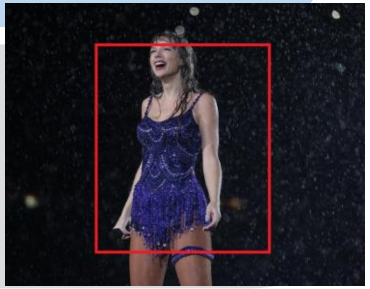
GPT-V The image shows a close-up of a person's foot wearing a white **Nike Air Force 1 sneaker**. The sneaker is a low-top style with a white rubber sole and white laces. …

TASK Object Counting

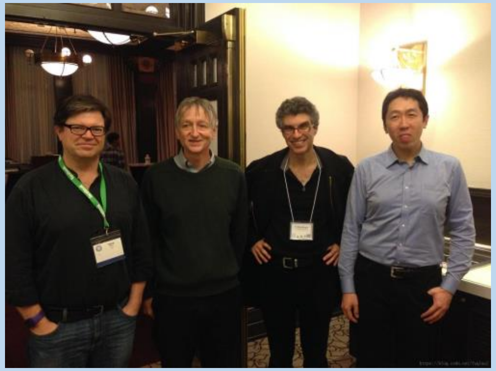
GPT-V There are **7 apples** in the image.

TASK Object Localization/Detection

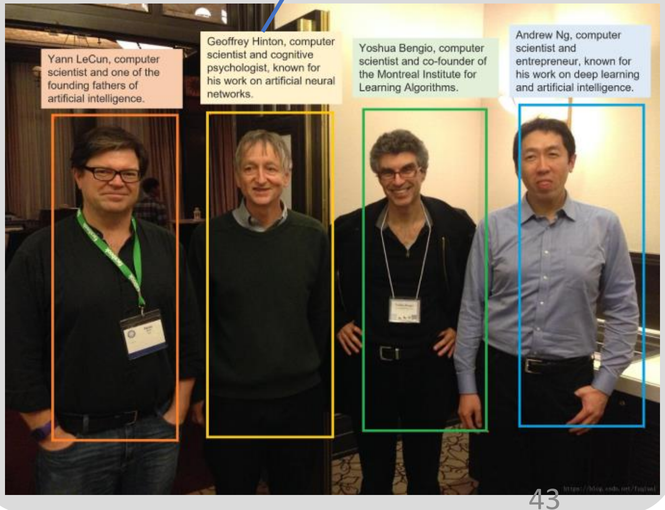
GPT-V There is only one person in the image, so the bounding box for this person would be:
- Top left: (254, 118)
- Bottom right: (731, 693)



TASK Dense Captioning



GPT-V **Geoffrey Hinton**, computer scientist and cognitive psychologist, known for his work on artificial neural networks.


Yann LeCun, computer scientist and one of the founding fathers of artificial intelligence.
Geoffrey Hinton, computer scientist and cognitive psychologist, known for his work on artificial neural networks.
Yoshua Bengio, computer scientist and co-founder of the Montreal Institute for Learning Algorithms.
Andrew Ng, computer scientist and entrepreneur, known for his work on deep learning and artificial intelligence.

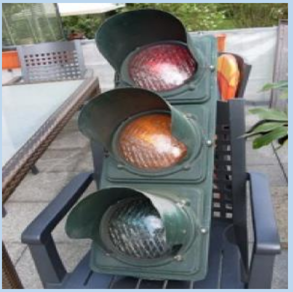
Yang, Z., et al. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv:2309.17421*, 2023.

Image credit to Lijuan Wang

What does GPT-V mean to computer vision?

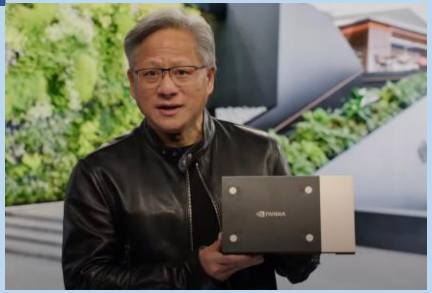
Microsoft

TASK **Image Recognition**



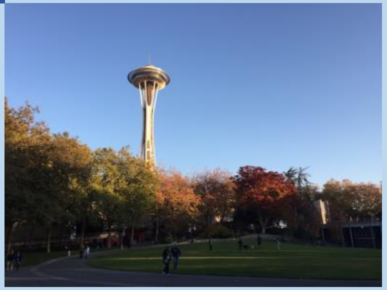
GPT-V The object in the image is a **traffic light**.

TASK **Celebrity Recognition**



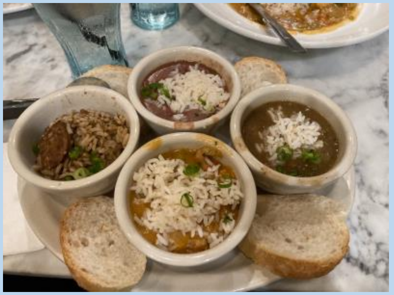
GPT-V The person in the image is **Jensen Huang**, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

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**

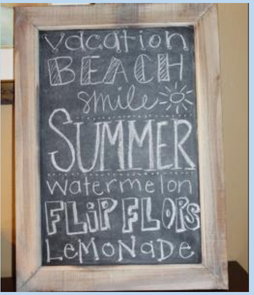


GPT-V This dish is called a **Cajun Sampler**. It is a sampler platter that consists of four small bowls of traditional Cajun dishes - gumbo, jambalaya, red beans and rice, and crawfish étouffée. …

Image credit to Lijuan Wang

What does GPT-V mean to computer vision?

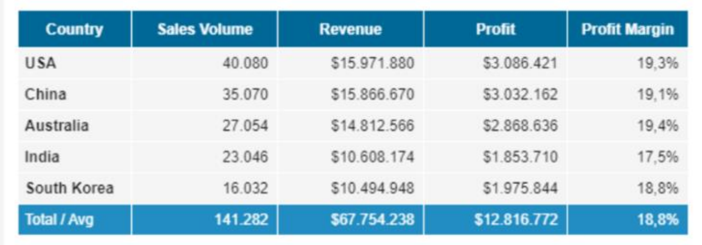
TASK **Scene Text Recognition**



GPT-V The scene text in the image are:
- Vacation - Beach - Smile
- Summer - Watermelon
- Flip flops - Lemonade

TASK **Document Reasoning**

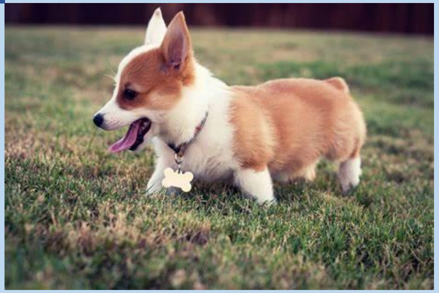
Which country has the highest profit margin?



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%

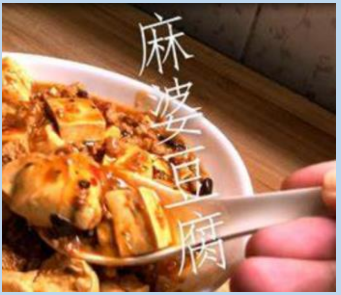
GPT-V Based on the data provided, the country with the highest profit margin is **Australia**, with a profit margin of 19,4%.

TASK **Multilingual 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 **Multilingual Text Recognition**

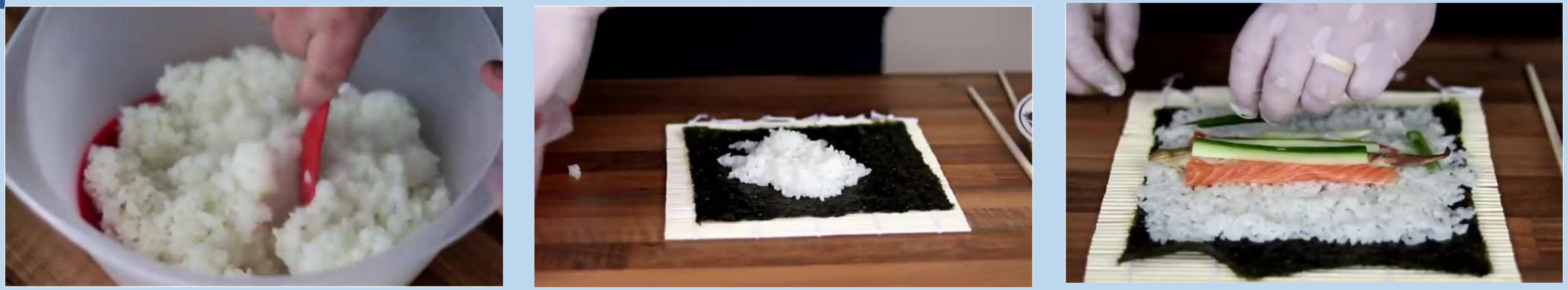


GPT-V The text in the image is in Chinese and translates to "**Mapo Tofu**." Mapo tofu is a popular Chinese dish made with tofu and minced meat in a spicy sauce.

Image credit to Lijuan Wang

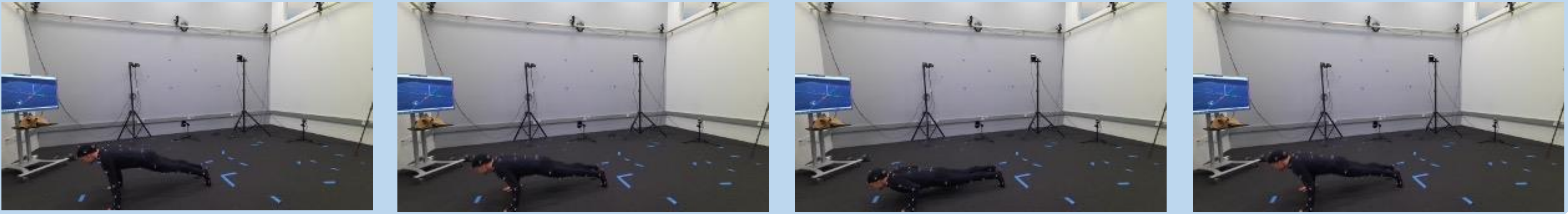
What does GPT-V mean to computer vision?

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**



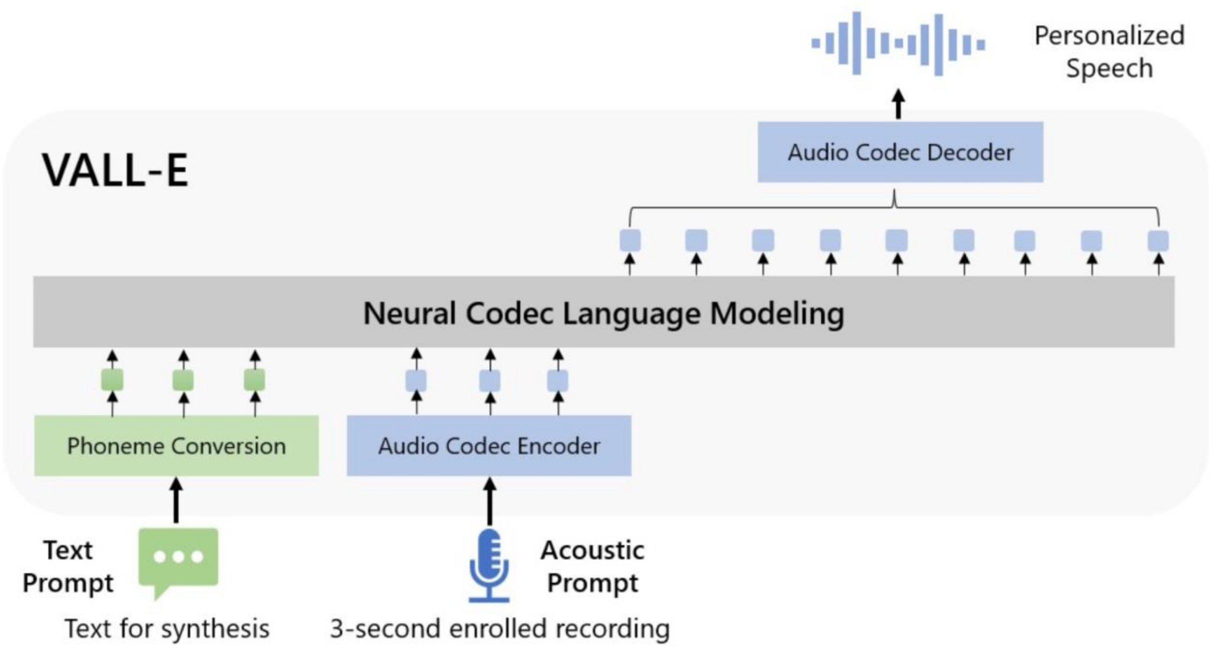
GPT-V The images show a person in a motion capture suit in a studio with multiple cameras and sensors. The person is **performing a push-up exercise**. In the first image, the person is in the starting position with their arms extended and body in a straight line. In the second image, the person is lowering 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.

Image credit to Lijuan Wang

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	🔈
🔈		🔈
🔈		🔈

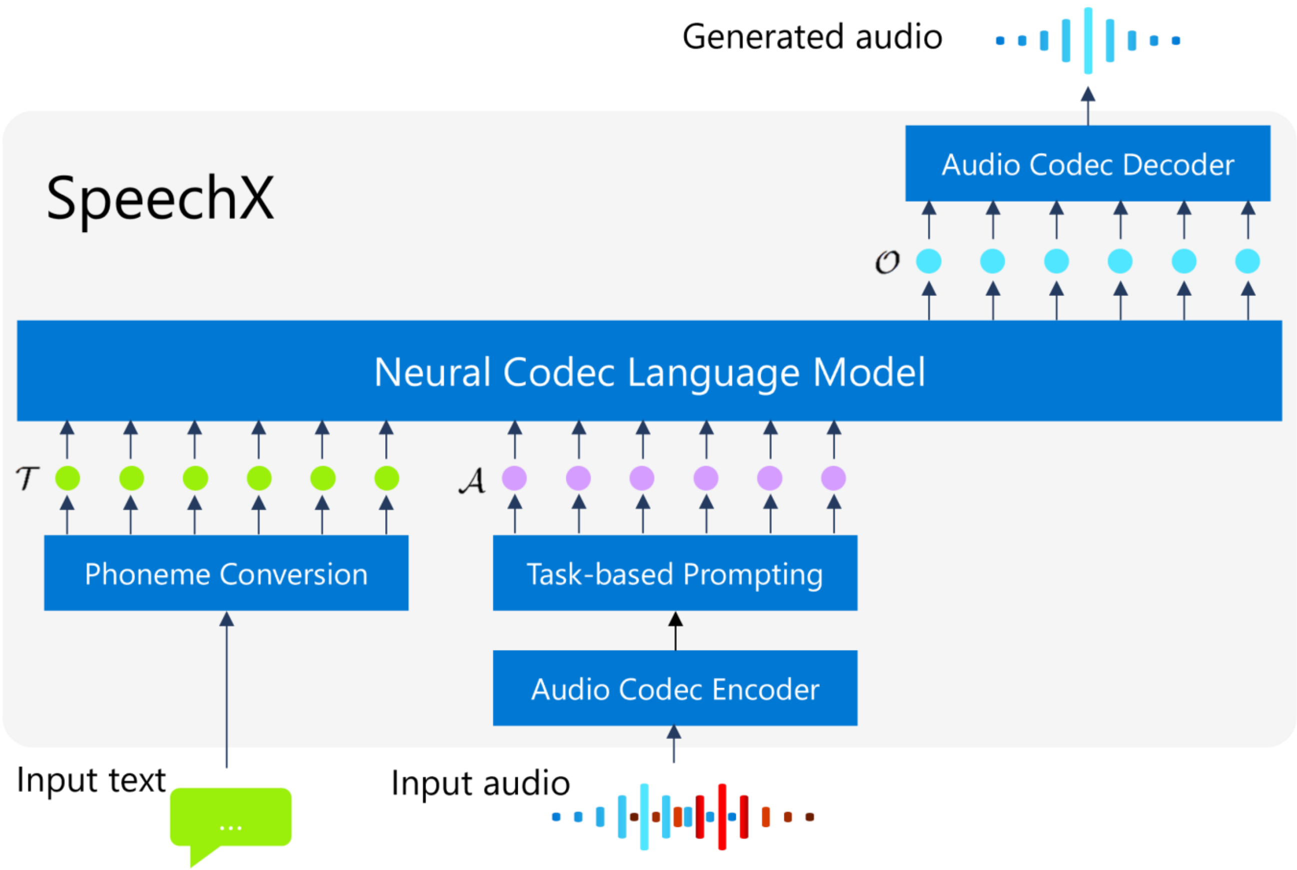
Wang, C., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv:2301.02111, 2023*.

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.

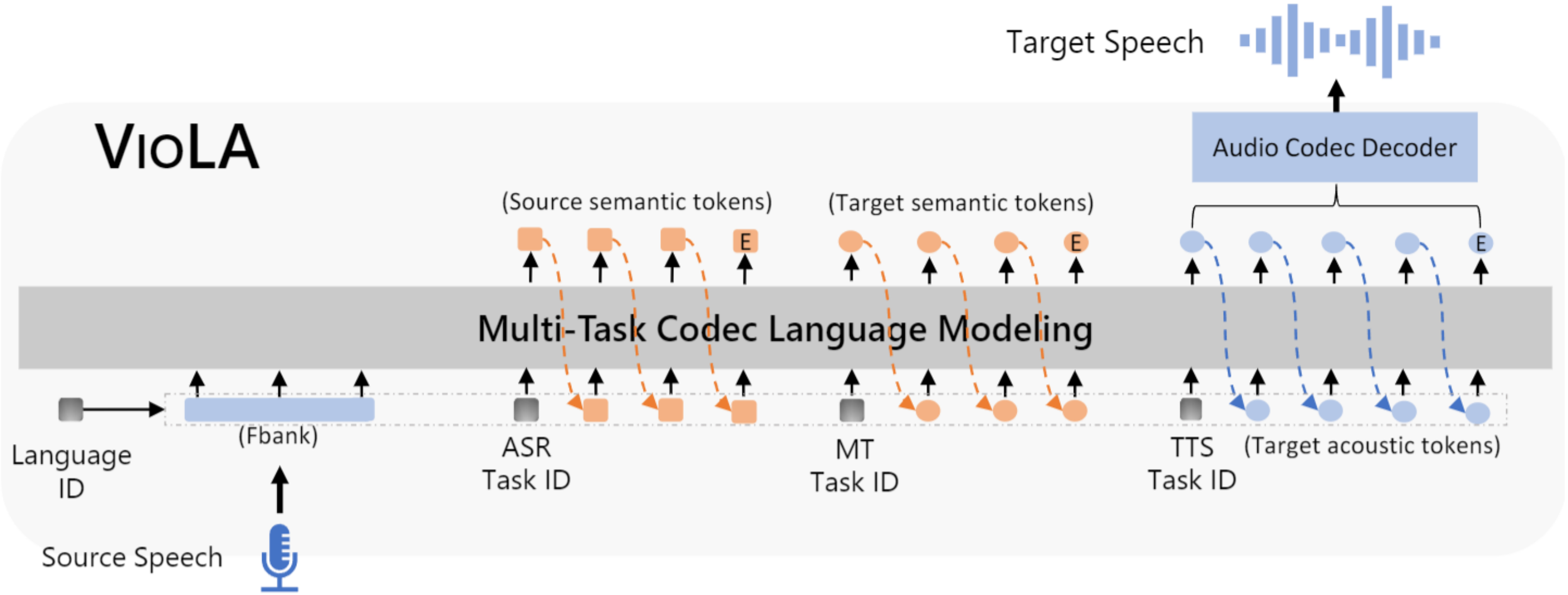


Task	Input text	Input audio	Output audio
Noise suppression	Transcription (optional)	Noisy speech 🔊	Clean speech 🔊
Speech removal	Transcription (optional)	Noisy speech 🔊	Noise 🔊
Target speaker extraction	Transcription (optional)	Speech mixture, Enrollment speech 🔊	Clean speech of target speaker 🔊
Zero-short 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]

Wang, X., et al. Speechx: Neural codec language model as a versatile speech transformer. *arXiv:2308.06873,* 2023.

VioLA: A multi-modal model with discrete audio inputs



Speech and text can freely serve as input and output

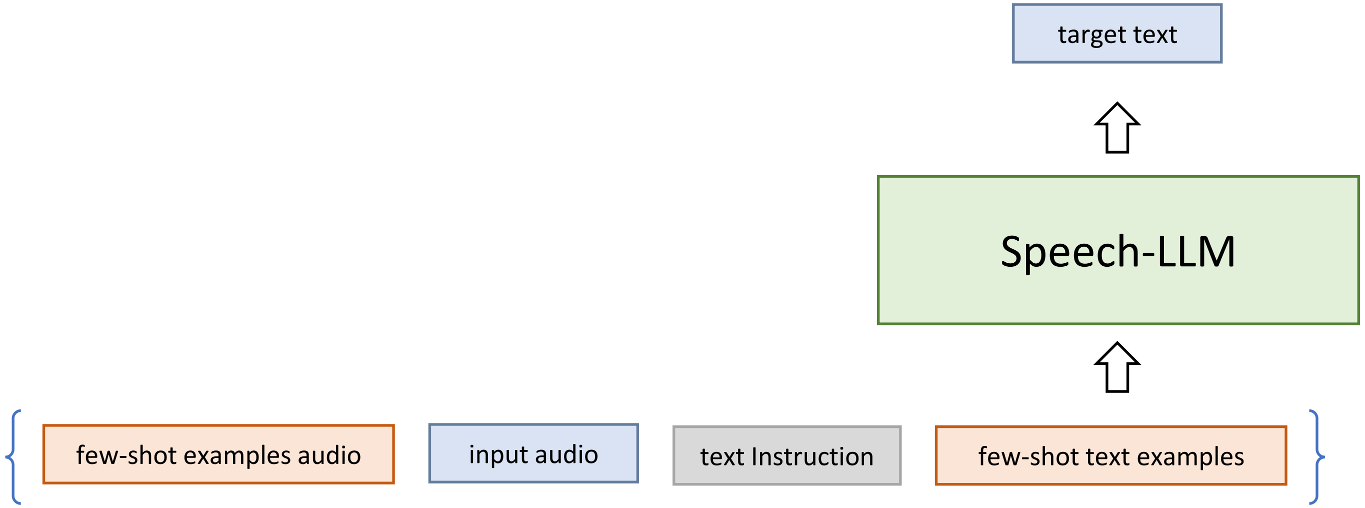
- An extension to audio codec language model
- Naturally merge speech-language tasks
  - Speech recognition
  - Machine translation
  - Speech generation

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

Wang, T., et al., VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation. *arXiv:2305.16107, 2023*.

Advancing Speech-LLM For In-context Learning

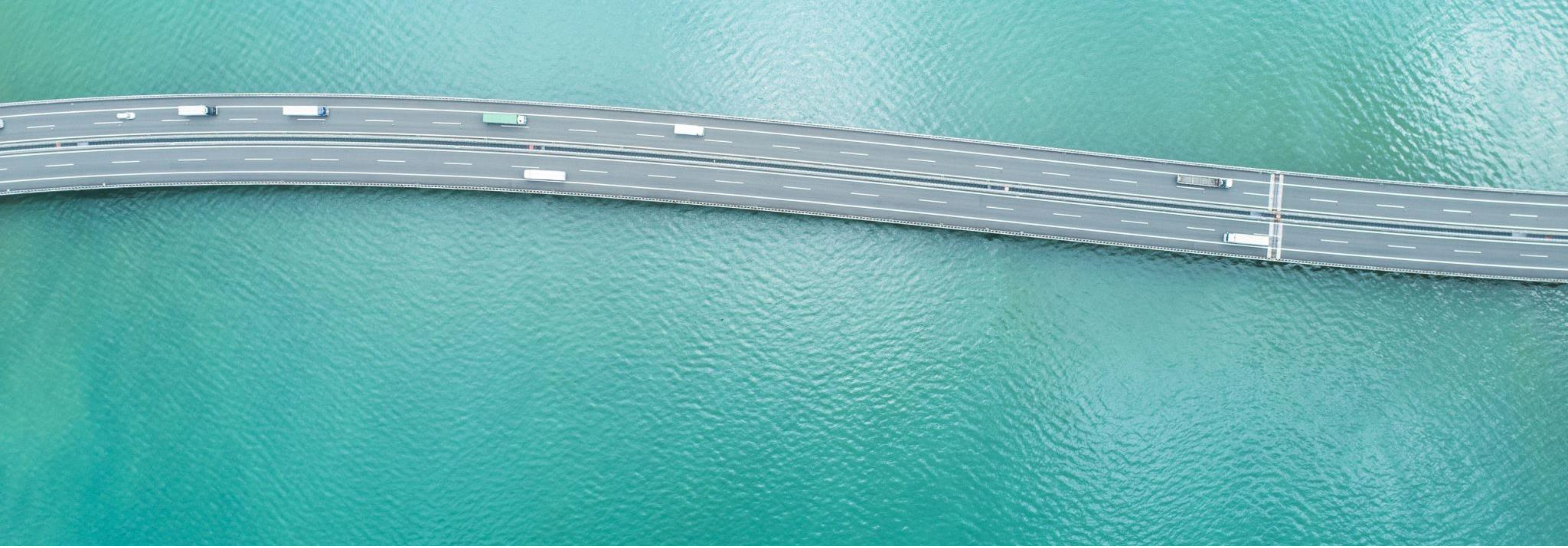
- Trained tasks (EN only)
  - ASR
  - Speech-based Question Answering
- Emergent Capable tasks
  - 0-shot and 1-shot En->X ST
  - 1-shot domain adaptation
  - Instruction-followed ASR



Wu, J., et al. On decoder-only architecture for speech-to-text and large language model integration. *In Proc. ASRU*, 2023.
Pan, J., et al. COSMIC: Data Efficient Instruction-tuning For Speech In-Context Learning. *arXiv preprint*, 2023.

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!