End-to-End
Automatic Speech Recognition
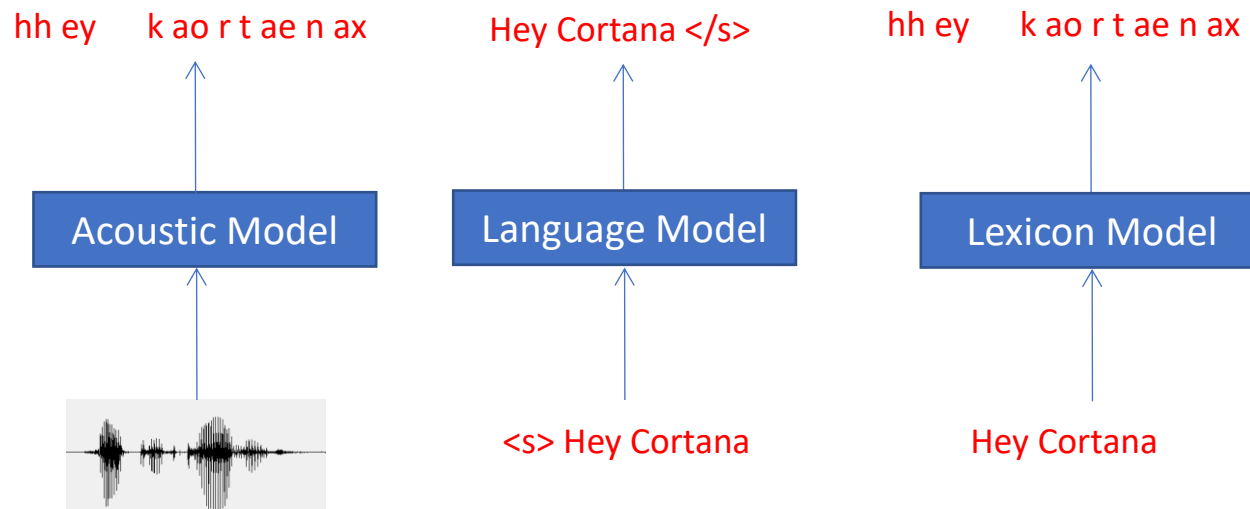
Jinyu Li

Microsoft

Outline

- End-to-end (E2E) automatic speech recognition (ASR) fundamental

- E2E advances
  - Leveraging unpaired text
  - Multi-talker ASR
  - Beyond ASR

- The next trend

- Conclusions
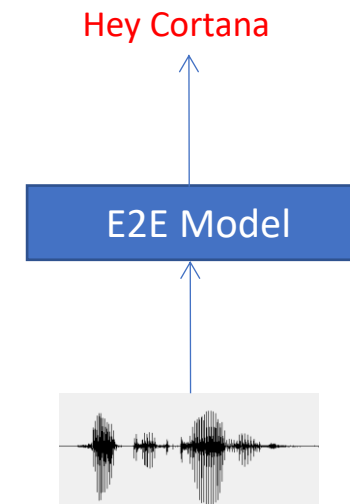
End-to-End Fundamental

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

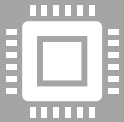A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks" PMLR, 2014.
A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," in arXiv preprint, 2014.

Current Status

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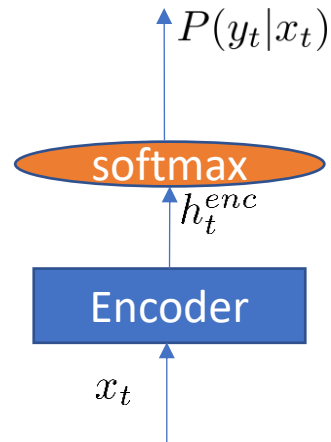
Practical challenges such as streaming, latency, adaptation capability etc., have been also optimized in E2E models.
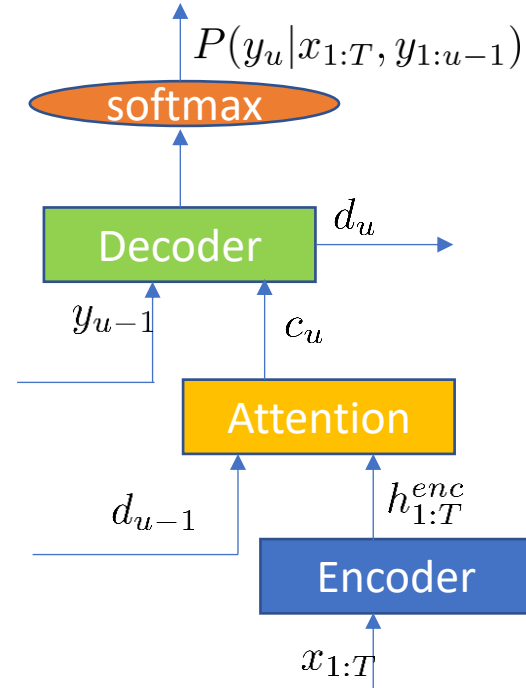
E2E models are now the mainstream models not only in academic but also in industry.
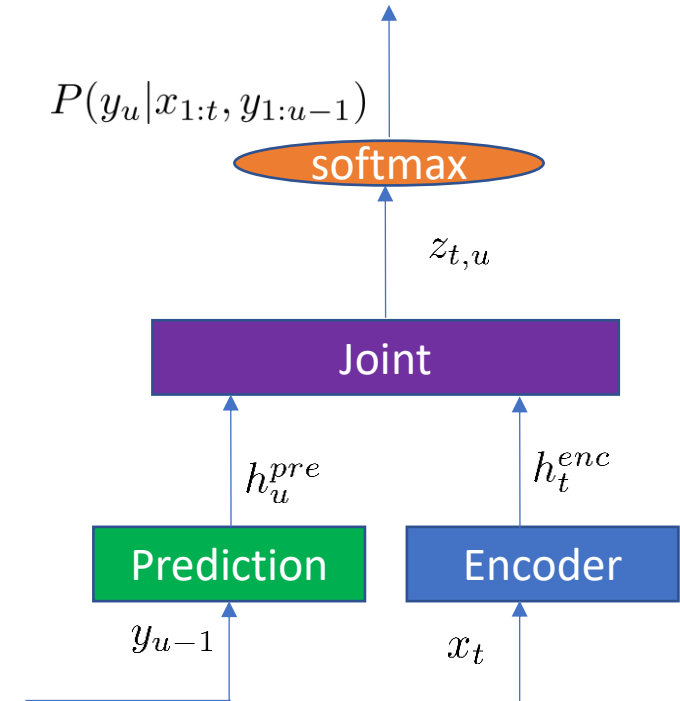
E2E Models

Connectionist Temporal Classification (CTC)

Attention-based encoder decoder (AED)
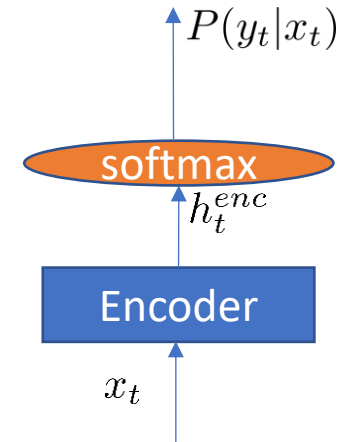
RNN-Transducer (RNN-T)

CTC

- The first and simplest E2E ASR model.
- To solve the challenge that target text label length is smaller than the speech input length:
  - Inserts blank and allows label repetition to have the same length of CTC path and speech input sequence.

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{q} \in \mathbf{B}^{-1}(\mathbf{y})} P(\mathbf{q}|\mathbf{x})$$

- Frame independence assumption

$$P(\mathbf{q}|\mathbf{x}) = \prod_{t=1}^{T} P(q_t|\mathbf{x})$$

- Revives with the Transformer encoder and the emerged self-supervised learning technologies

$P(y_t|x_t)$

softmax

h_t^{enc}

Encoder

x_t

A. Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proc. ICML, 2006.
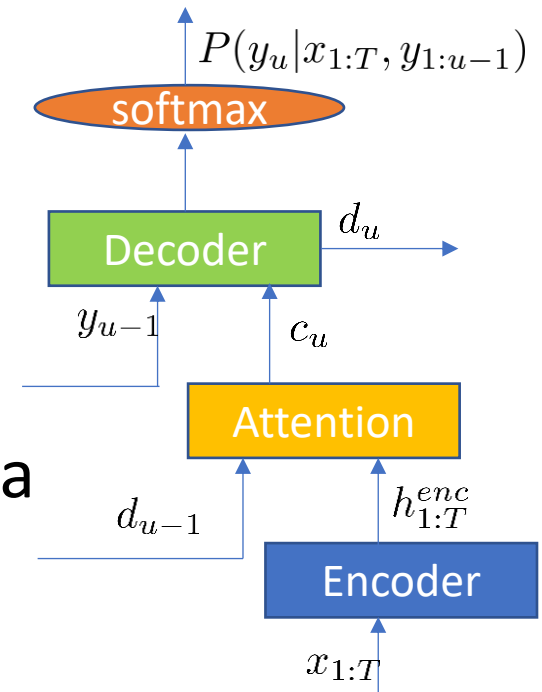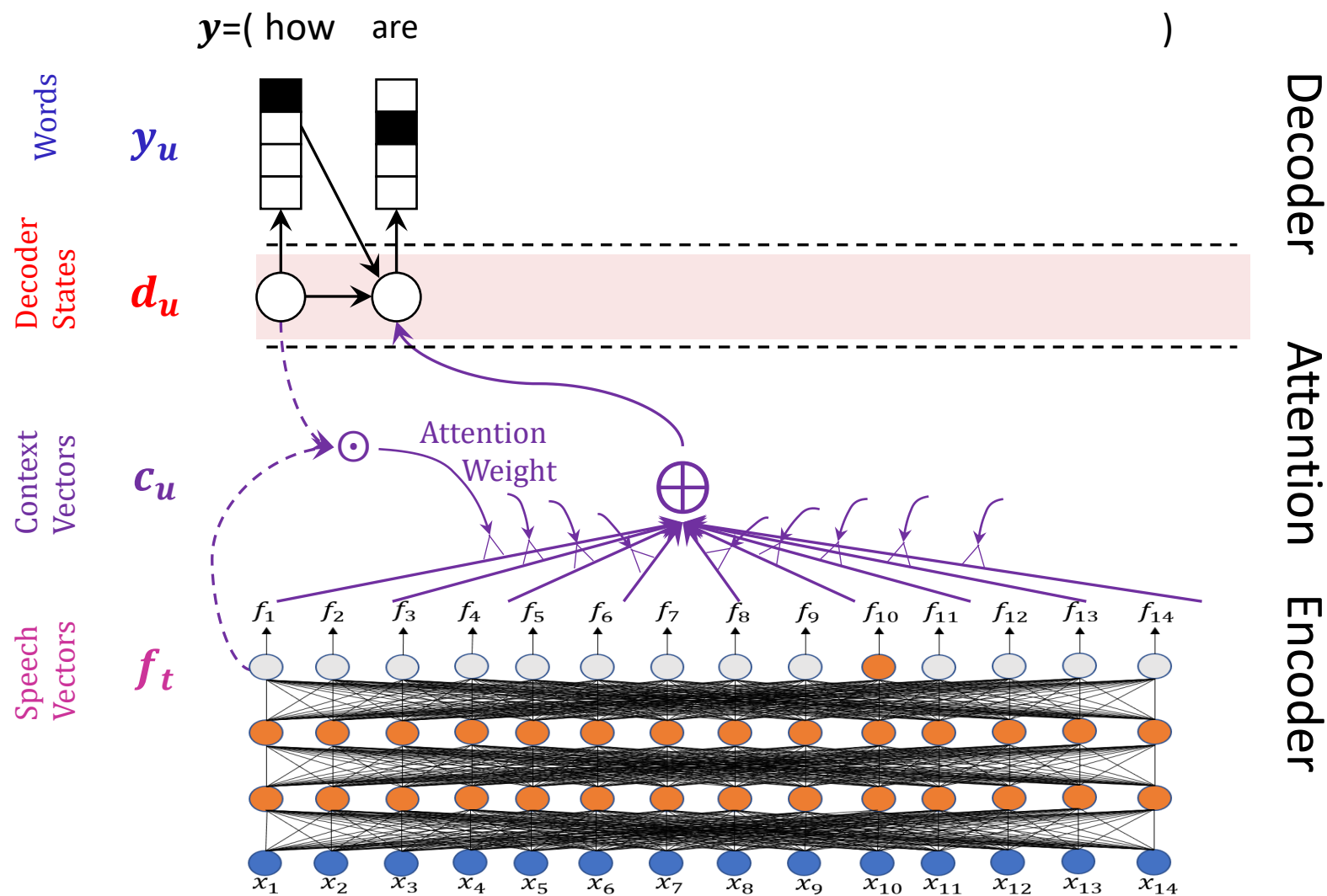
AED

- The sequence probability is calculated in an auto-regressive way.

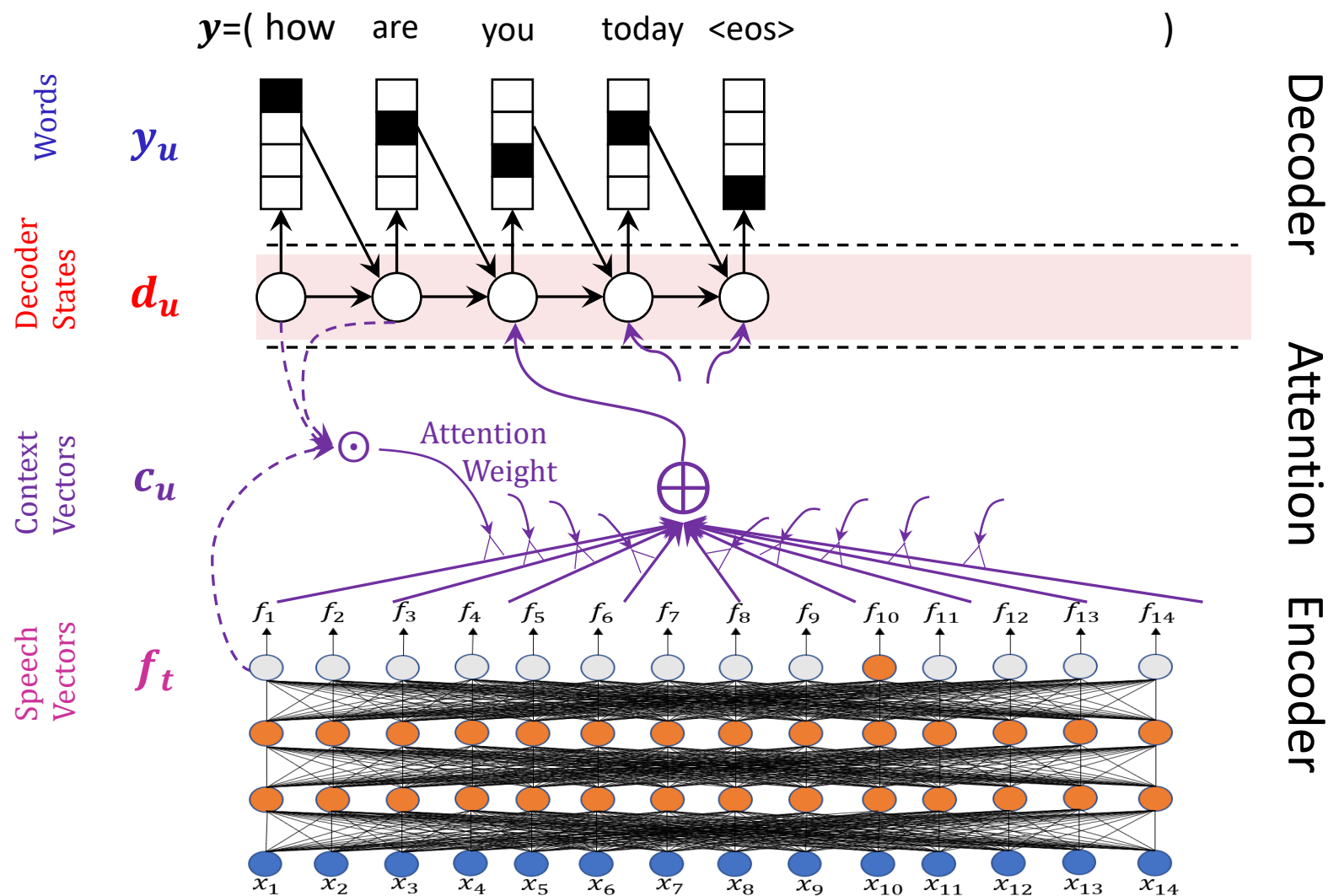$$P(\mathbf{y}|\mathbf{x}) = \prod_u P(y_u|\mathbf{x}, \mathbf{y}_{1:u-1})$$

- Encoder: converts input acoustic sequences into high-level hidden feature sequences.

- Attention: computes attention weights to generate a context vector as a weighted sum of the encoder output.

- Decoder: takes the previous output label together with the context vector to generate its output $P(y_u|\mathbf{x}, \mathbf{y}_{1:u-1})$

$P(y_u|x_{1:T}, y_{1:u-1})$

softmax

Decoder

d_u

y_{u-1}

c_u

Attention

d_{u-1}

$h^{enc}_{1:T}$

Encoder

$x_{1:T}$

J. Chorowski, et al., "End-to-end continuous speech recognition using attention-based recurrent NN: First results," arXiv:1412.1602, 2014.
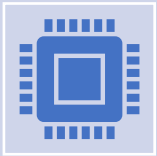
AED Inference

AED Inference

Streaming

Lots of commercial setups need the ASR systems to be streaming with low latency: ASR system produces the recognition results at the same time as the user is speaking.

Non-streaming ASR is not practical in lots of ASR scenarios where speech signal comes in a continuous mode without segmentation.
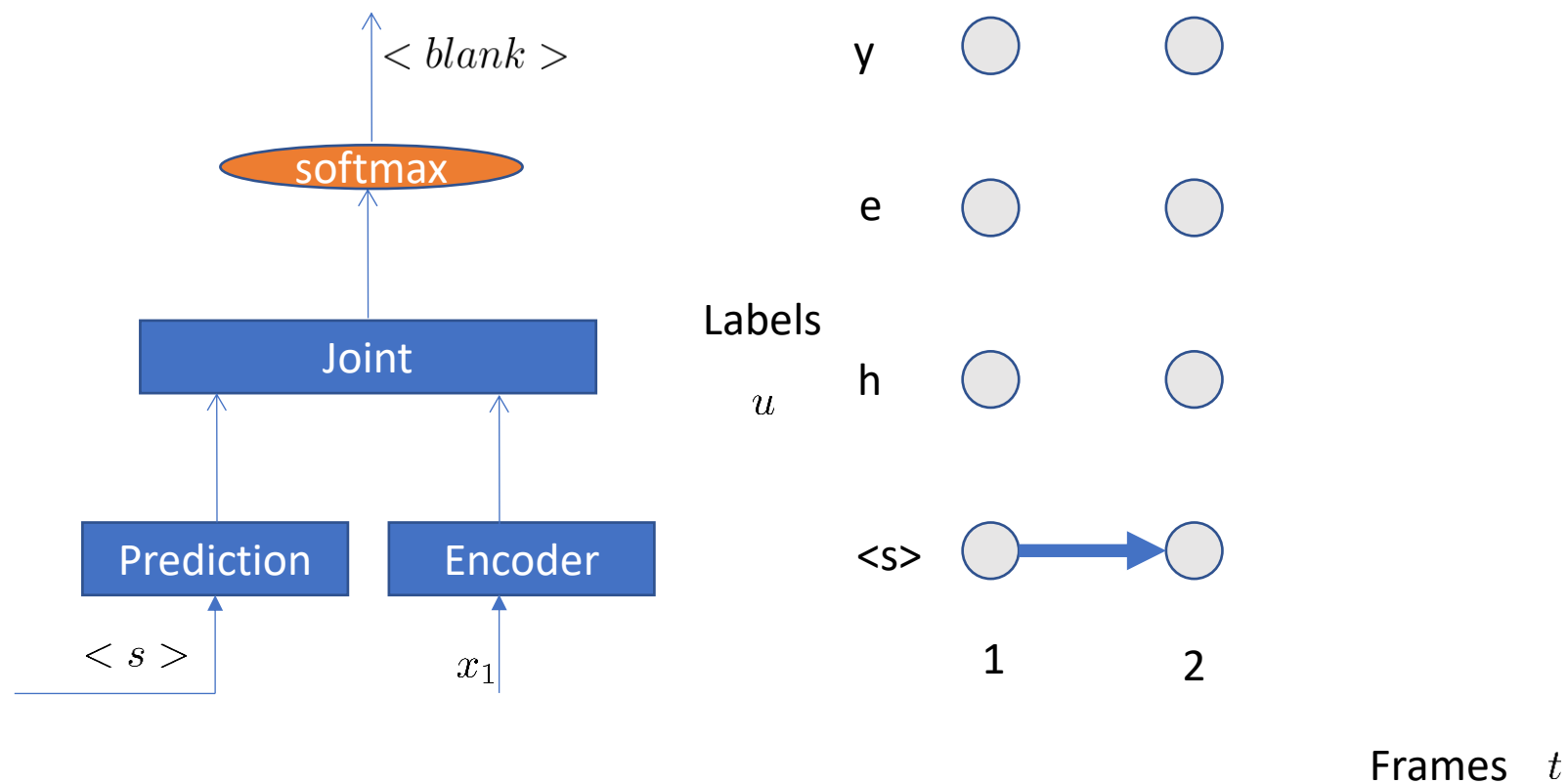
Streaming

- Full attention in AED cannot work for streaming ASR
  - Streaming AED (MOCHA, MILK etc.): apply attention on chunks of input speech.
  - Not a natural design for streaming.

- RNN-T provides a natural way for streaming ASR and becomes the most popular E2E model in industry.

C. Chiu and C. Raffel, "Monotonic chunkwise attention," in Proc. ICLR, 2018.
N. Arivazhagan et al., "Monotonic infinite lookback attention for simultaneous machine translation," in Proc. ACL, 2019.
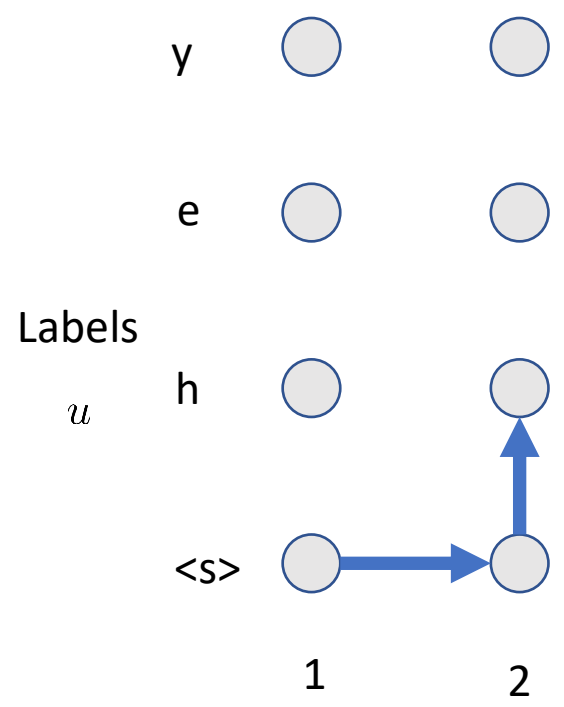
RNN-T Path

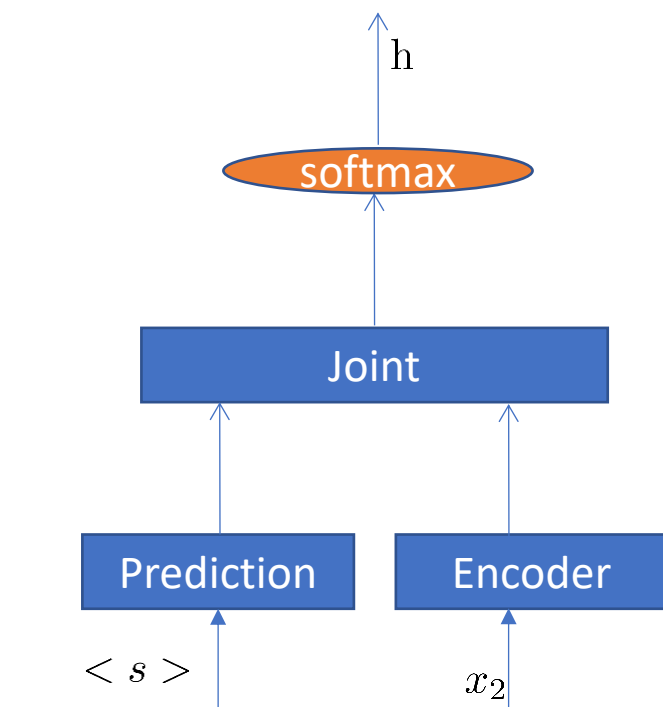$< blank >$ output: advance encoder, otherwise, advance prediction network.

A. Graves., "Sequence transduction with recurrent neural networks," in arXiv preprint, 2012.

RNN-T Path

$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path

$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path

$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path

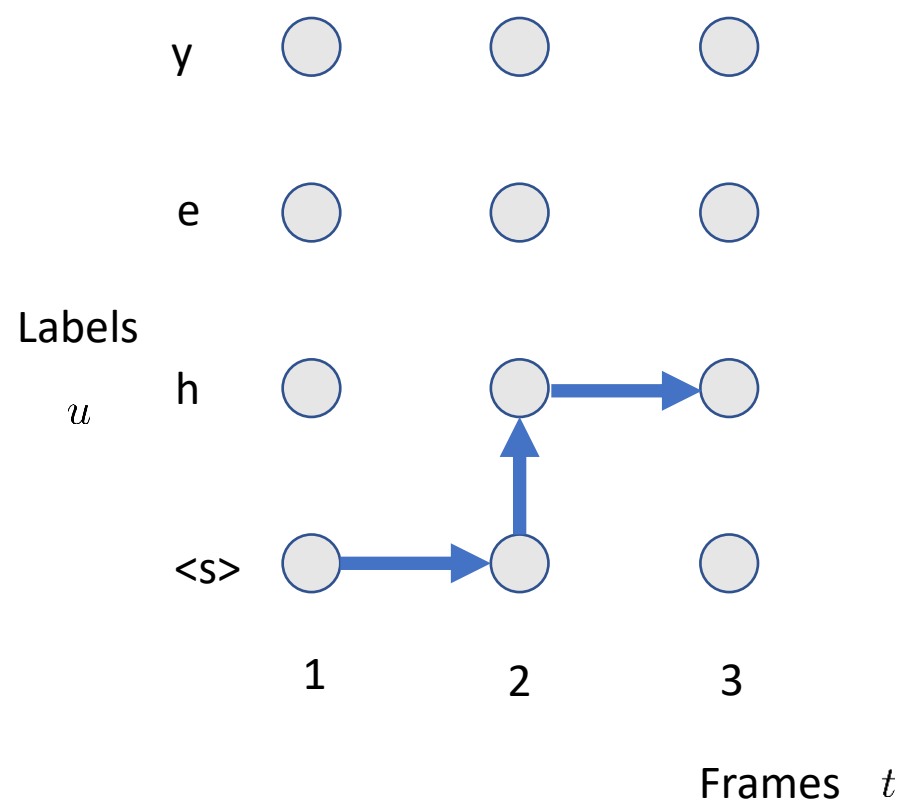$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path

RNN-T Path

$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path

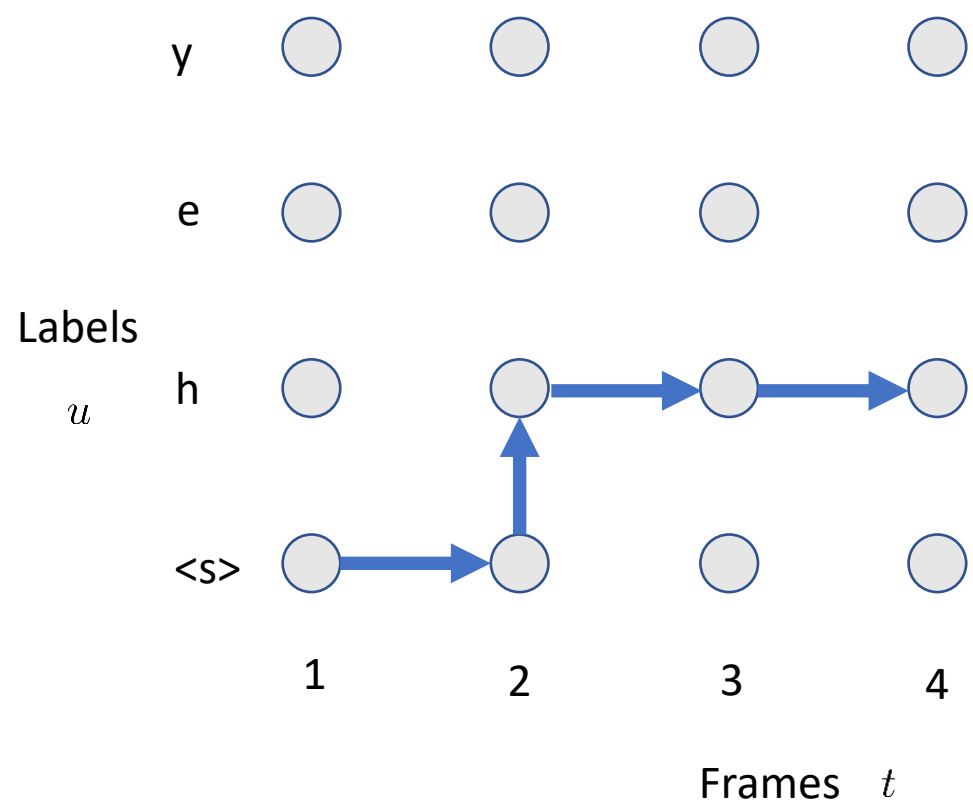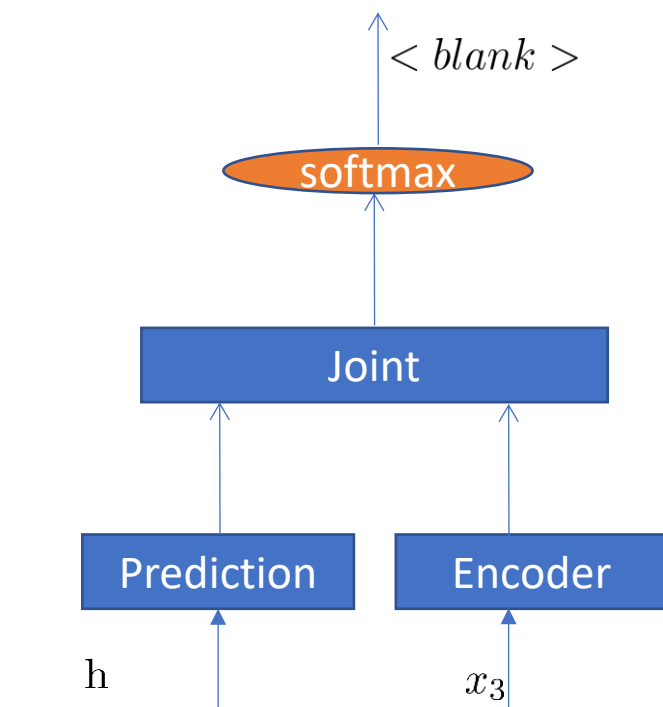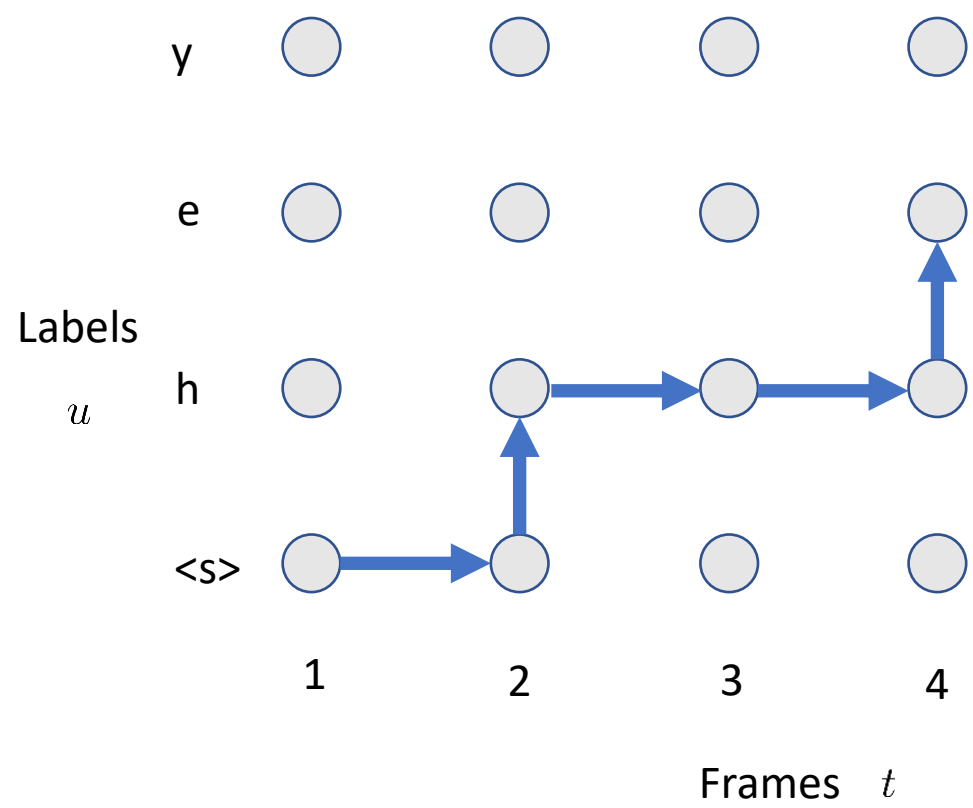$< blank >$ output: advance encoder, otherwise, advance prediction network.

RNN-T Path
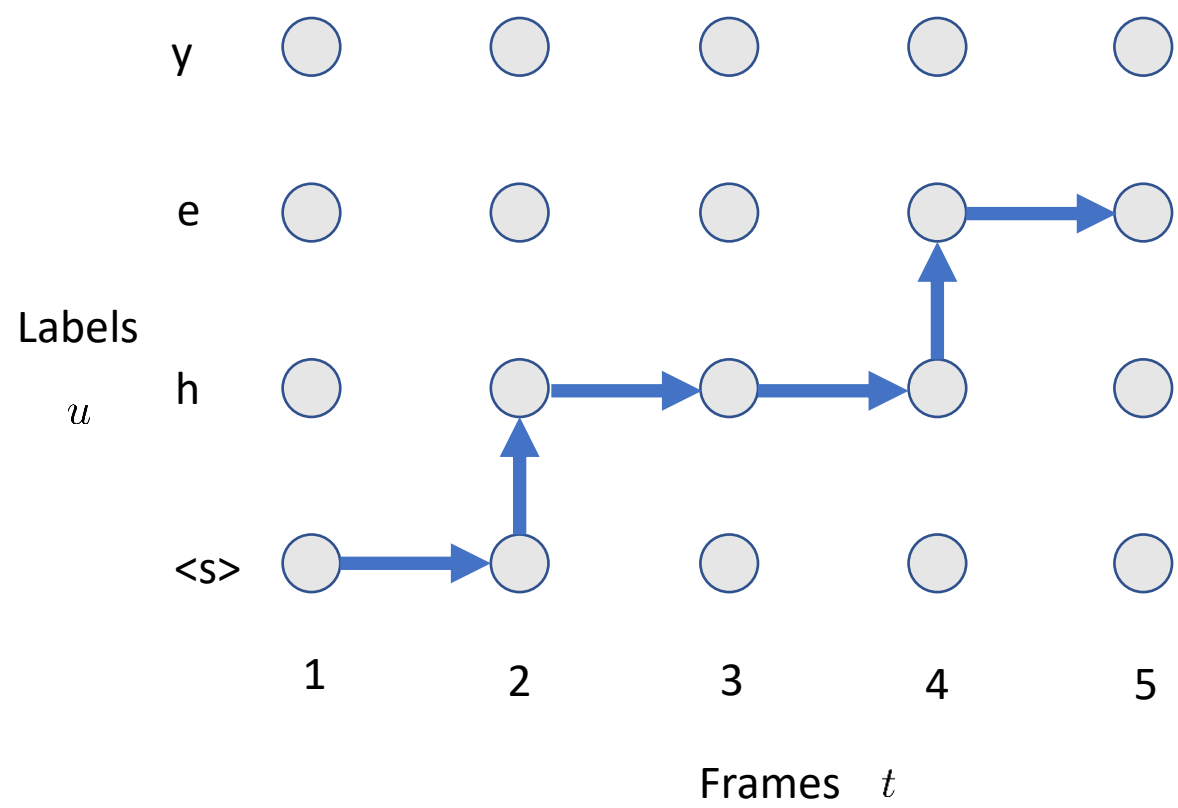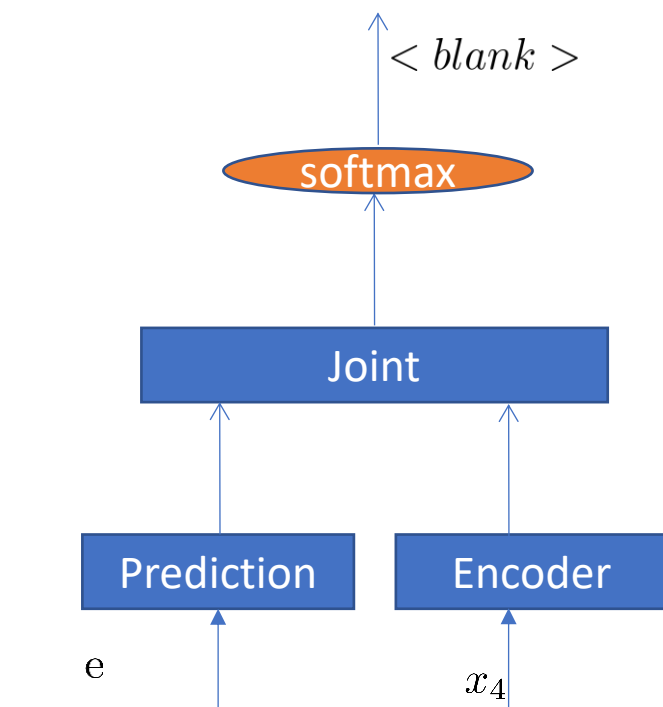
RNN-T Path

RNN-T Training

Given a label sequence of length U and acoustic frames of length T, the training maximizes the probabilities of all RNN-T paths.
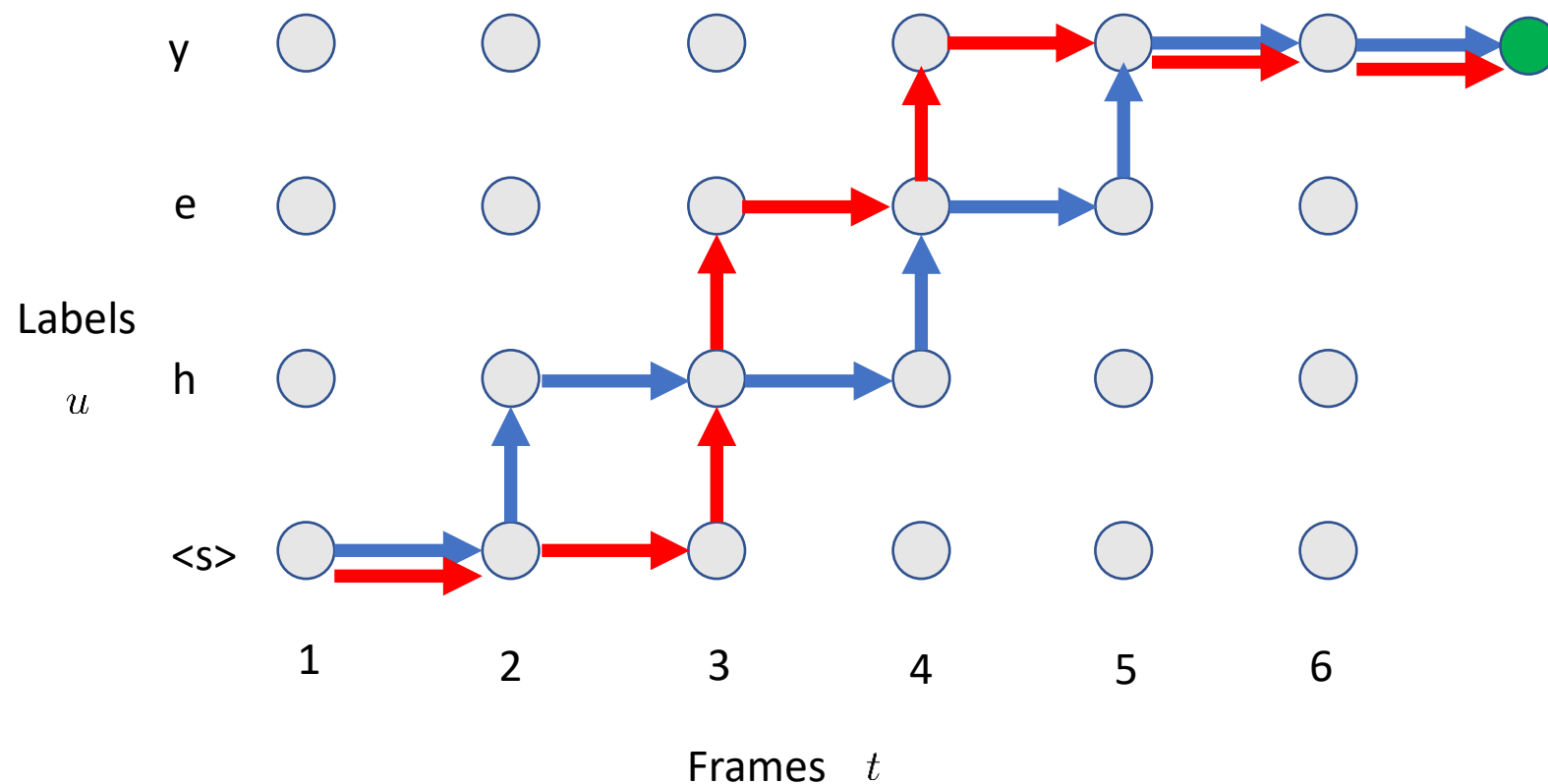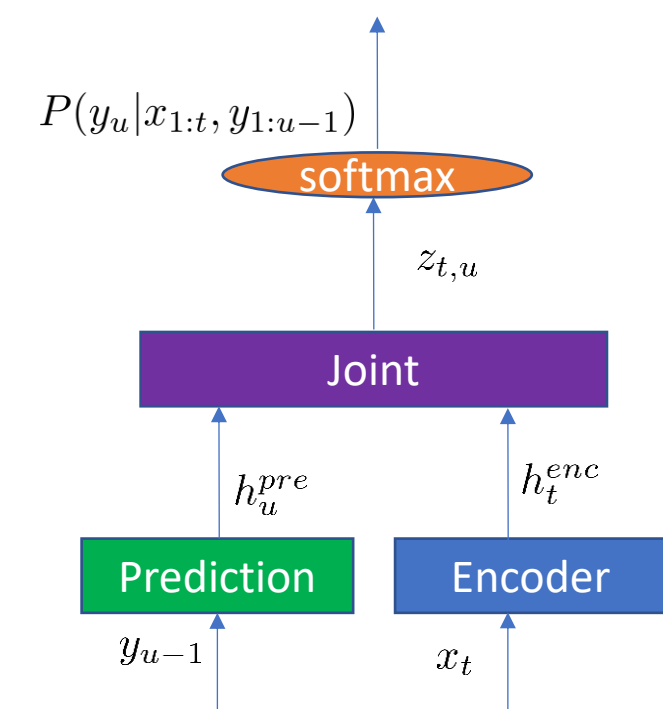
E2E Models

	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

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

J. Li, et al., "Developing RNN-T models surpassing high-performance hybrid models with customization capability," in Proc. Interspeech, 2020.
T. Sainath, et al., "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in Proc. ICASSP, 2020.

Encoder is the Most Important Component
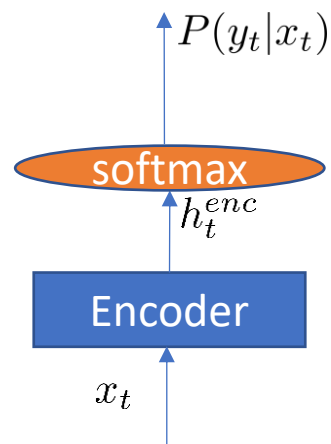
Connectionist Temporal Classification (CTC)
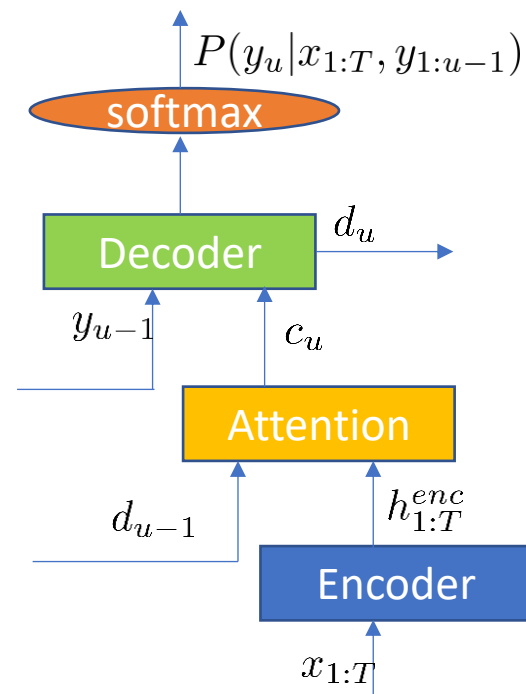
Attention-based encoder decoder (AED)
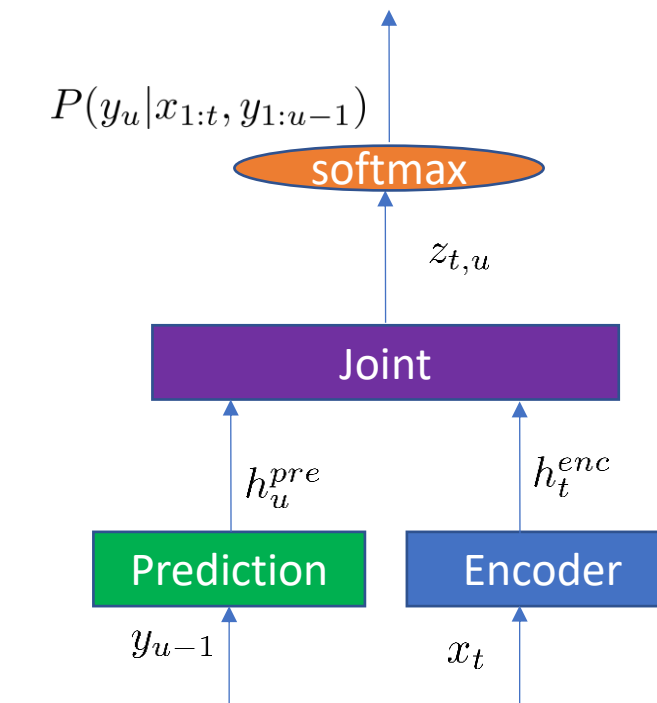
RNN-Transducer (RNN-T)

Encoder for RNN-T

$$P(y_u|x_{1:t}, y_{1:u-1})$$

softmax

$$z_{t,u}$$

Joint

$$h_u^{pre} \qquad h_t^{enc}$$

Prediction   Encoder

LSTM

Transformer

Conformer

$$y_{u-1} \qquad x_t$$
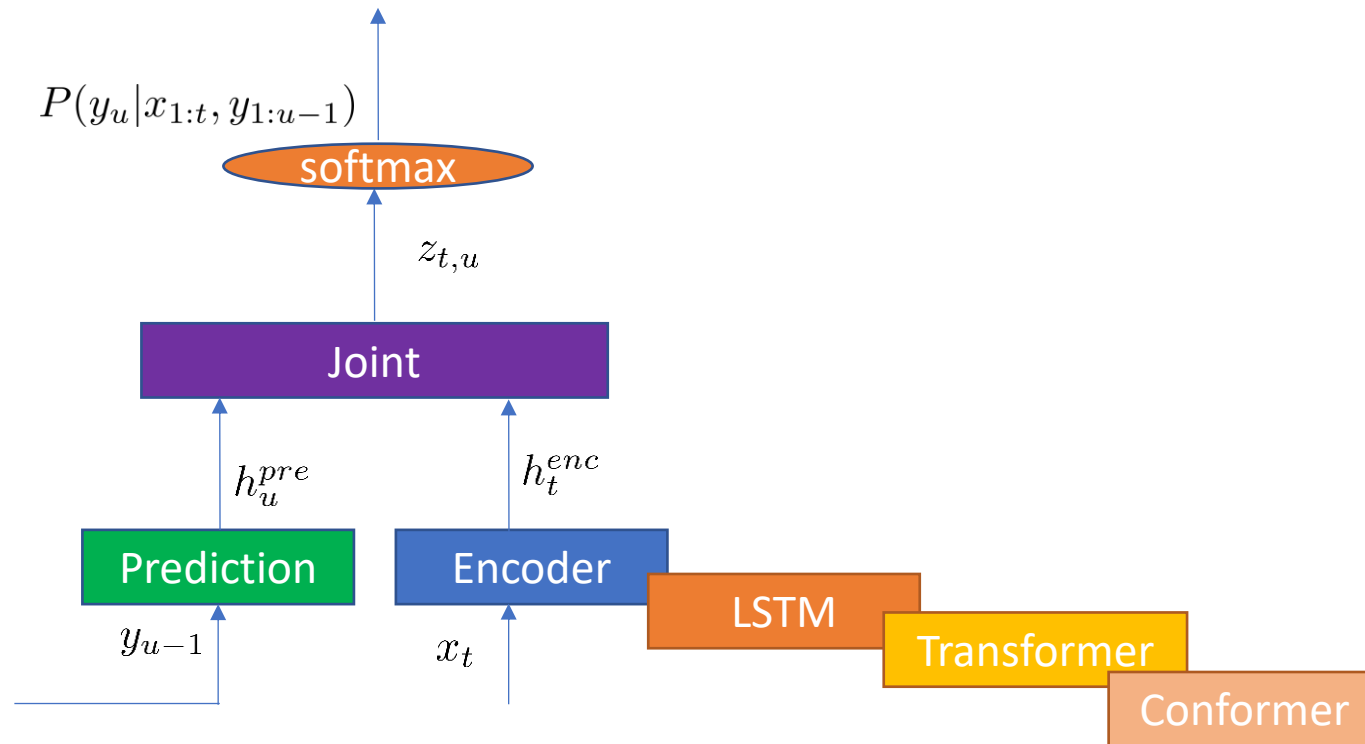
Transformer

- Self-attention: computes the attention weights 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

Input

A. Vaswani, et al., "Attention is all you need," NIPS, 2017.
Q. Zhang, et al., "Transformer transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss," in Proc. ICASSP, 2020.
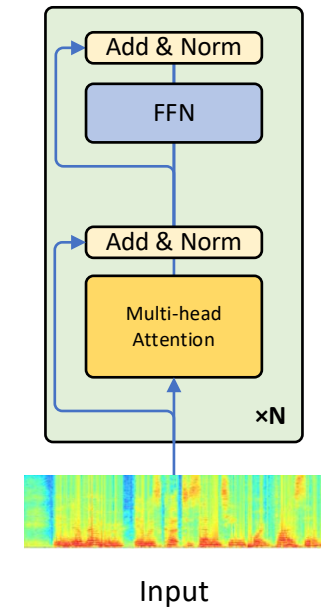
Conformer

- 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
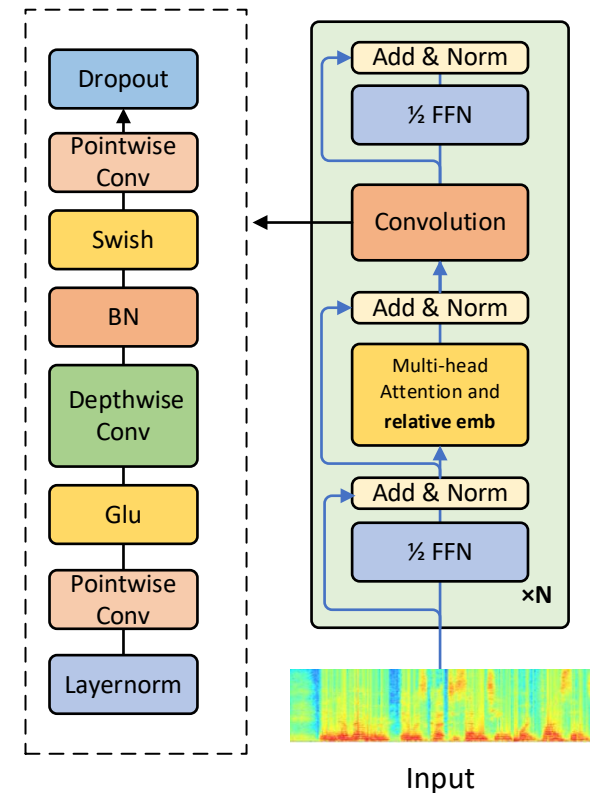
Input

A. 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

- In order to build streaming ASR, we need both the model and its encoder to be streaming.

- 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

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

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

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

X. Chen, 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)

generating output for x_{10} · **Not streamable** · Attention Mask

Attention Mask is All You Need

- 0 lookahead, full history

generating output for x_{10}

Memory and runtime cost increase linearly

Attention Mask

Attention Mask is All You Need

- 0 lookahead, limited history (3 frames)

generating output for x_{10}
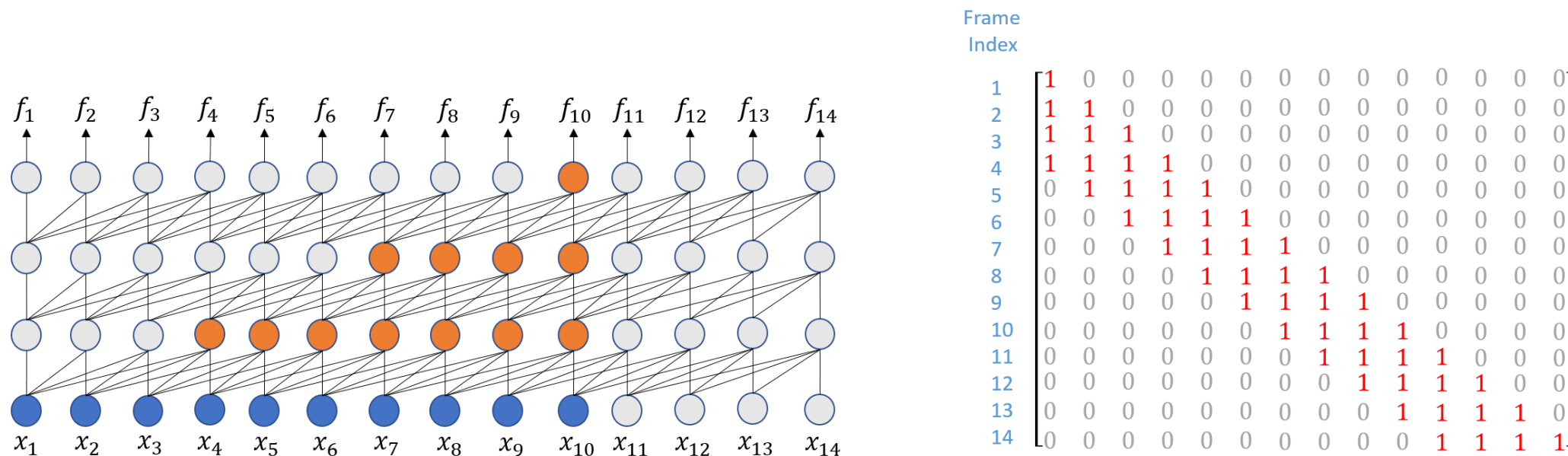
Attention Mask

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)

generating output for x_{10}

Look-ahead window [0, 2]

Attention Mask
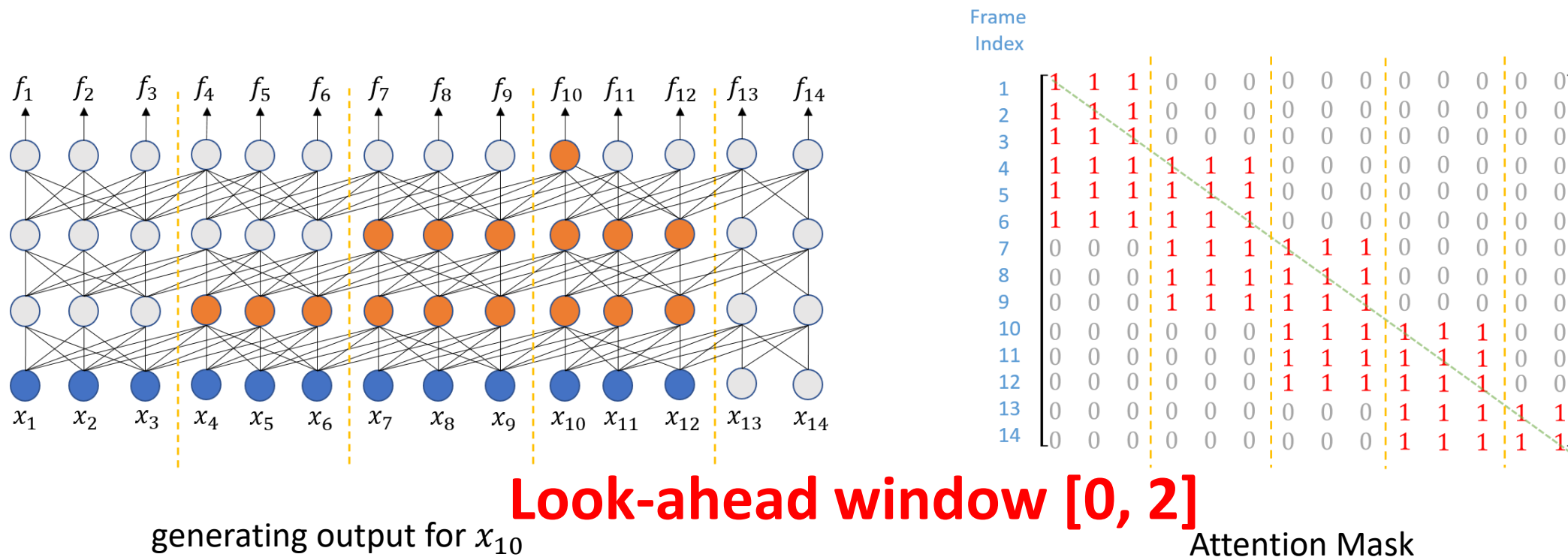
Live Caption in Windows 11
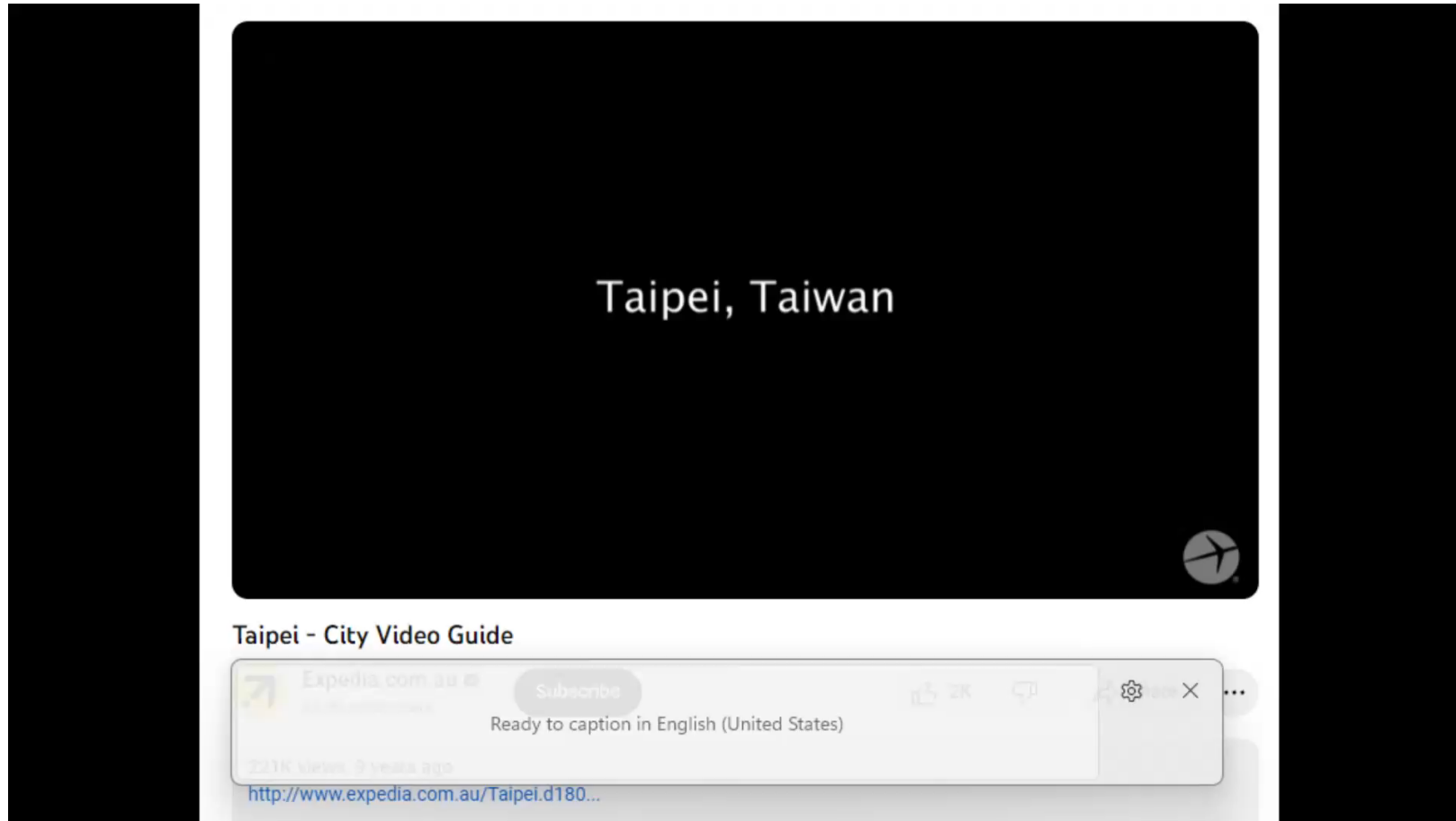
Advancing E2E Models
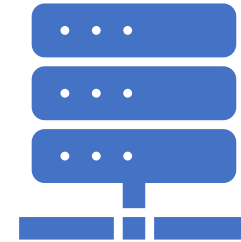
unpaired text   multi-talker ASR   beyond ASR

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 language model (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 either using TTS or from original ASR training data.
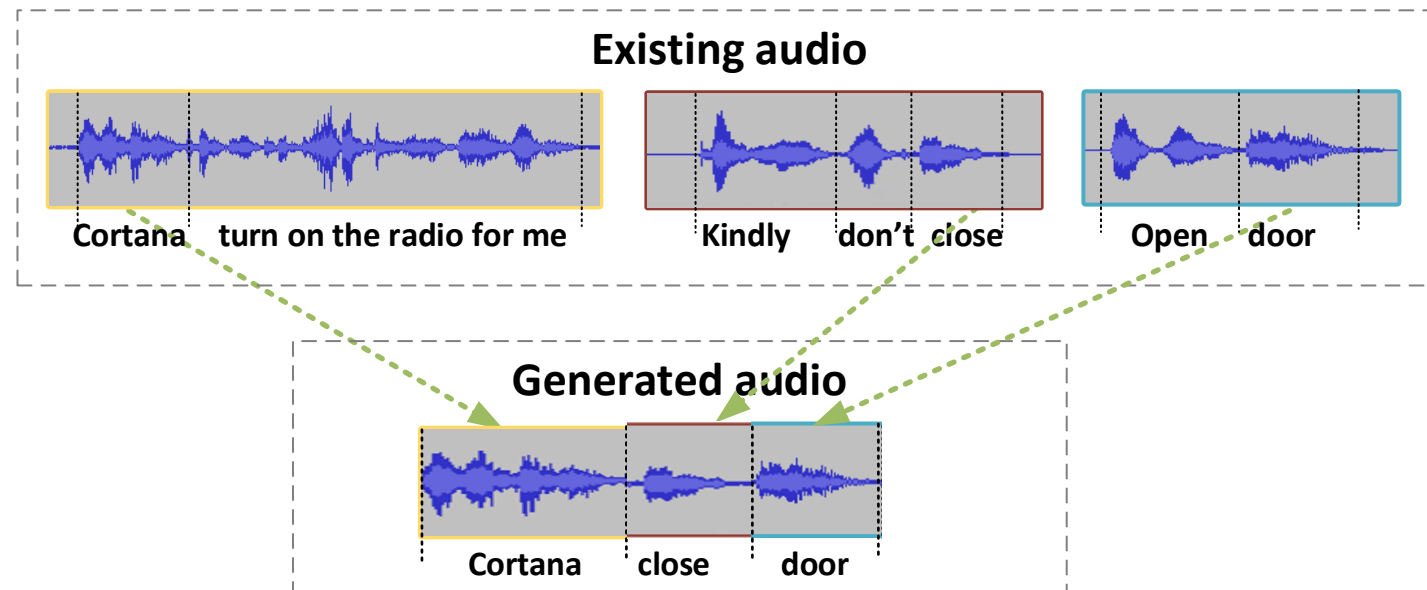
K. Sim, et al, "Personalization of end-to-end speech recognition on mobile devices for named entities," *in Proc. ASRU,* 2019.

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

R. Zhao, 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.

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

E2E score **Target LM score**

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

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

A standalone LM trained with training transcript of E2E model

Shallow Fusion score **Source LM score**

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

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

An inherent LM estimated from E2E model parameters

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

Shallow Fusion score **Internal LM score**

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

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

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

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

Internal LM Estimation

- **RNN-T**

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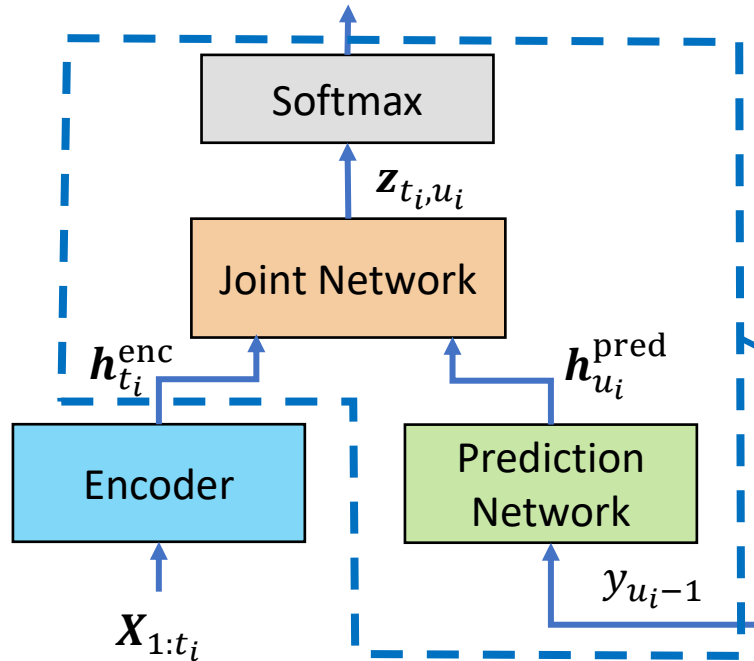
- **Internal LM estimation of RNN-T**

$$P(y_u | \boldsymbol{Y}_{0:u-1}; \theta_{\mathrm{pred}}, \theta_{\mathrm{joint}}) = \mathrm{softmax}(\boldsymbol{z}_u^{\mathrm{ILM,NB}})$$
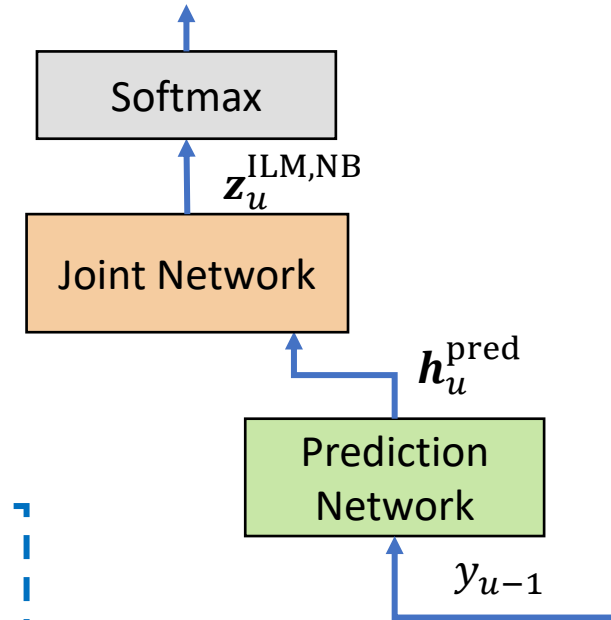
- Internal LM probability
  - The output of the **acoustically-conditioned LM** after removing the contribution of the encoder

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

Is the Prediction Network a LM?

- If the prediction network in RNN-T is a LM, we can use new-domain text to adapt it without even bothering audio data generation.

- However, it does not fully function as a LM because it needs to predict both vocabulary tokens and blank.

$P(y_u | x_{1:t}, y_{1:u-1})$

softmax

$z_{t,u}$

Joint

h_u^{pre} h_t^{enc}

Prediction   Encoder

y_{u-1} x_t

M. Ghodsi et al., "RNN-Transducer with stateless prediction network," *in Proc. ICASSP,* 2020.
X. Chen et al., "Factorized neural transducer for efficient language model adaptation," *in Proc. ICASSP,* 2022.

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!

X. Chen et al., "Factorized neural transducer for efficient language model adaptation," *in Proc. ICASSP,* 2022.

Multi-talker ASR

Multi-talker Models

- 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
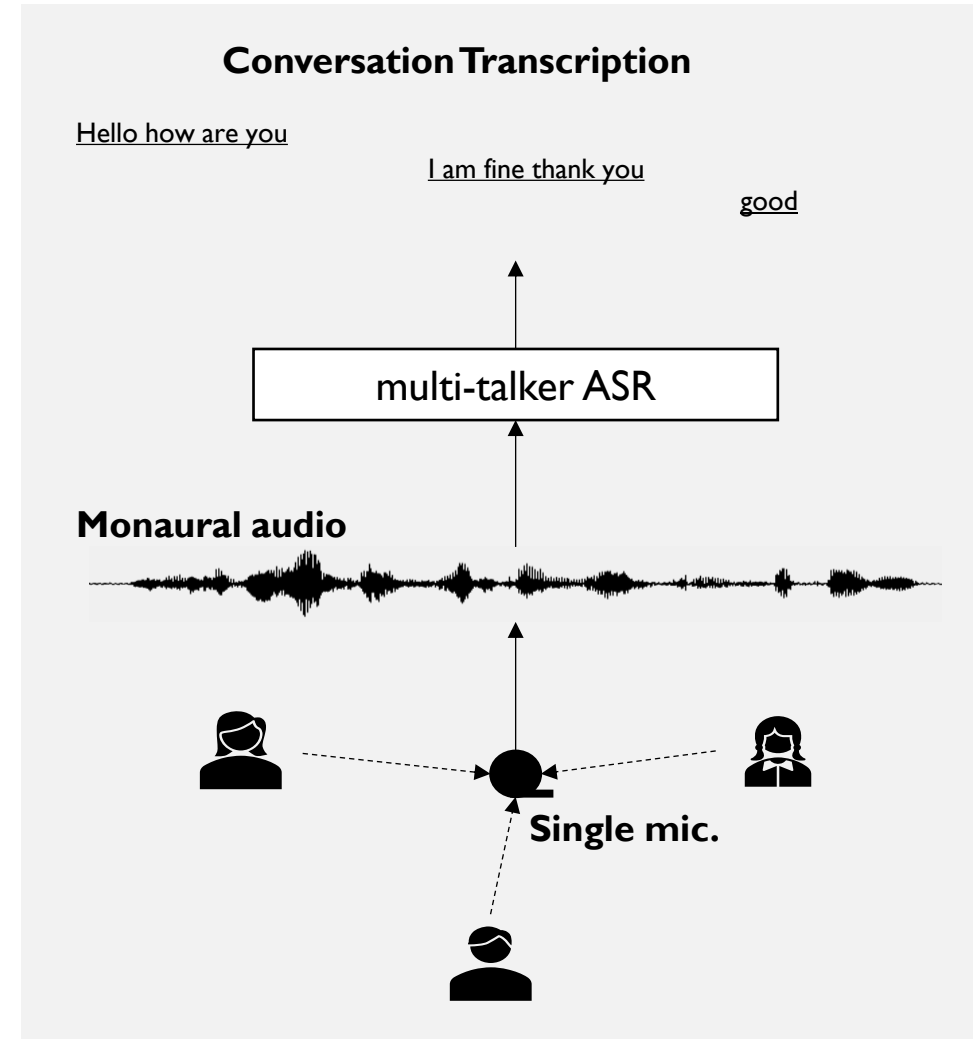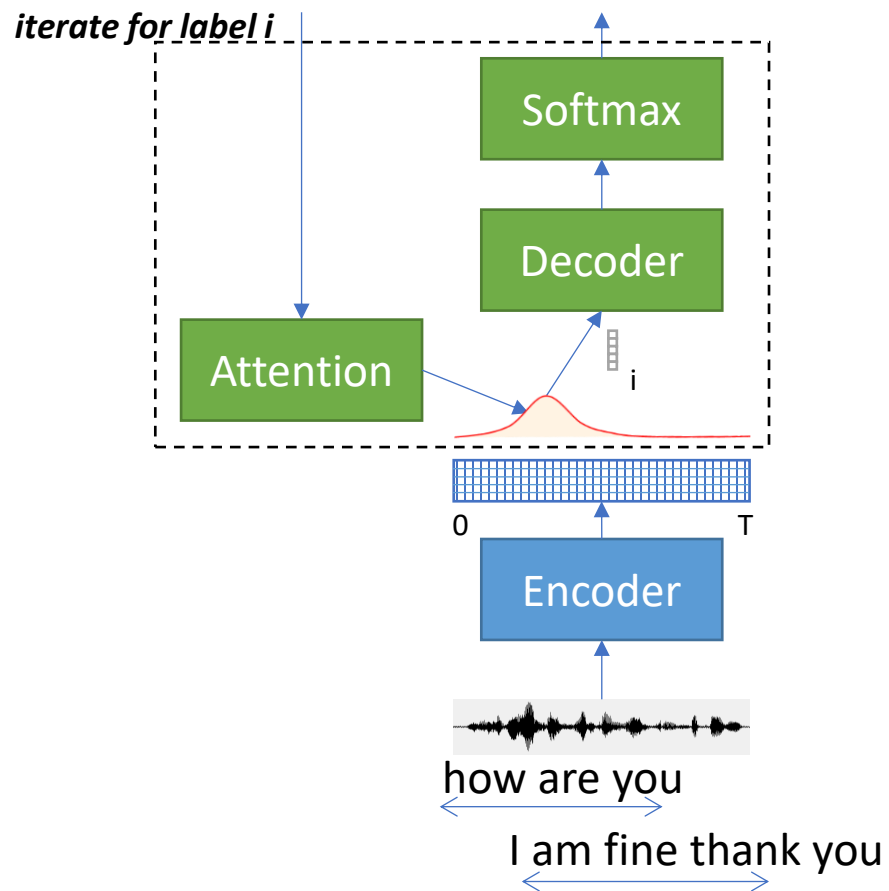
Conversation Transcription

Hello how are you

I am fine thank you

good

multi-talker ASR

Monaural audio

Single mic.

Serialized Output Training (SOT)

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

N. Kanda, et al., "Serialized output training for end-to-end overlapped speech recognition," In Proc. Interspeech, 2020.

Serialized Output Training (SOT)

how are you **<sc>** I am fine thank you **<eos>**
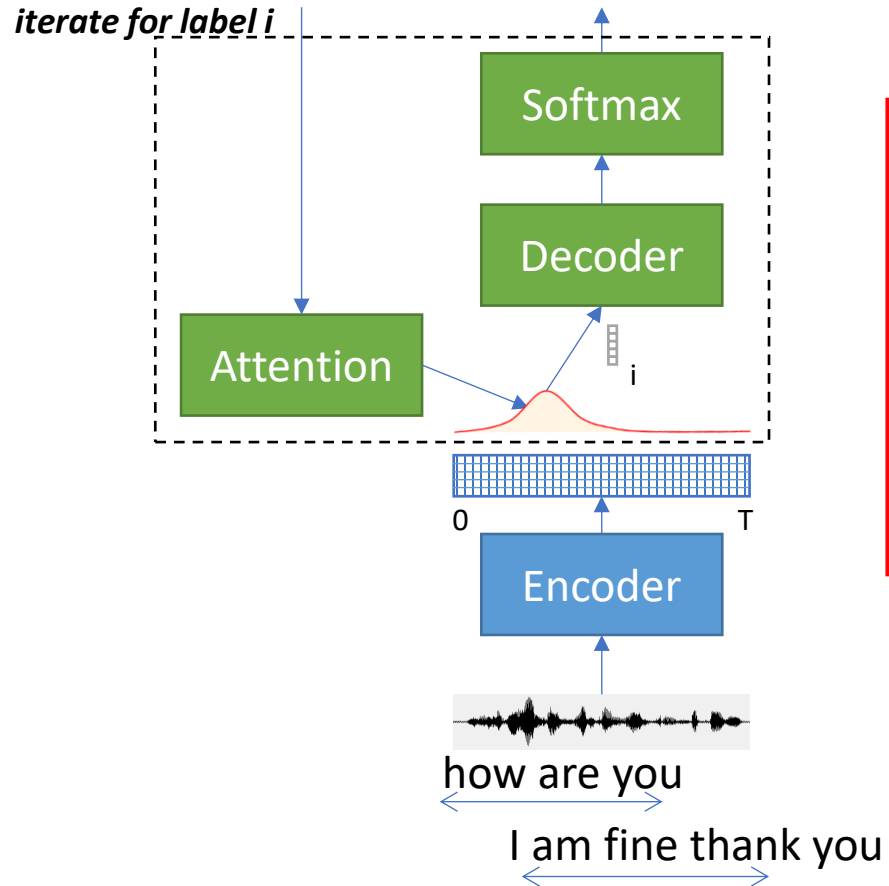
- Can recognize any number of speakers
- Achieved SOTA WERs for LibriSpeexhMix, LibriCSS, AMI, AliMeeting

Serialized Output Training (SOT)
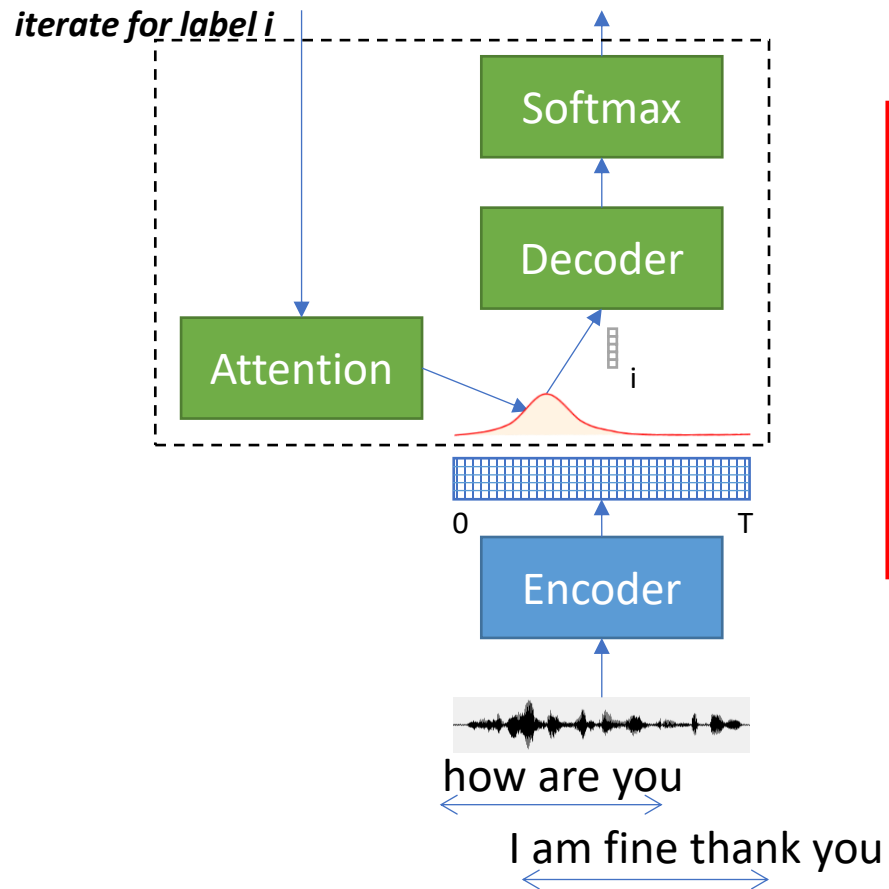
how are you <sc> I am fine thank you <eos>

iterate for label i

Softmax

Decoder

Attention

h_i

0   T

Encoder

how are you

I am fine thank you

- ☻
  - Can recognize any number of speakers
  - Achieved SOTA WERs for LibriSpeexhMix, LibriCSS, AMI, AliMeeting

- ☹ Only applicable for attention-based encoder decoder architecture
  → Only applicable for offline (i.e. non-streaming) inference

Token-level Serialized Output Training (t-SOT)

N. Kanda, et al., "Streaming speaker-attributed ASR with token-level speaker embeddings." in Proc. Interspeech, 2022.

Token-level Serialized Output Training (t-SOT)

Token-level Serialized Output Training (t-SOT)

Token-level Serialized Output Training (t-SOT)

Token-level Serialized Output Training (t-SOT)

t-SOT vs. SOT

- t-SOT is streamable
- t-SOT can be used for any type of ASR architecture
- t-SOT has limit on max concurrent utterances

SOT
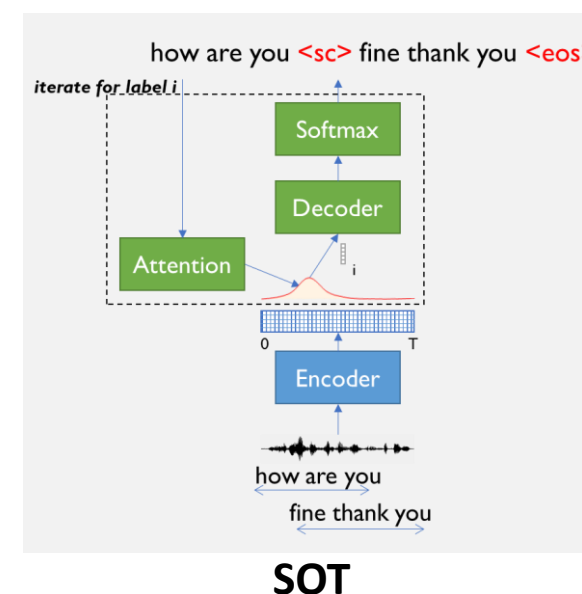
Multi-talker ASR Demo
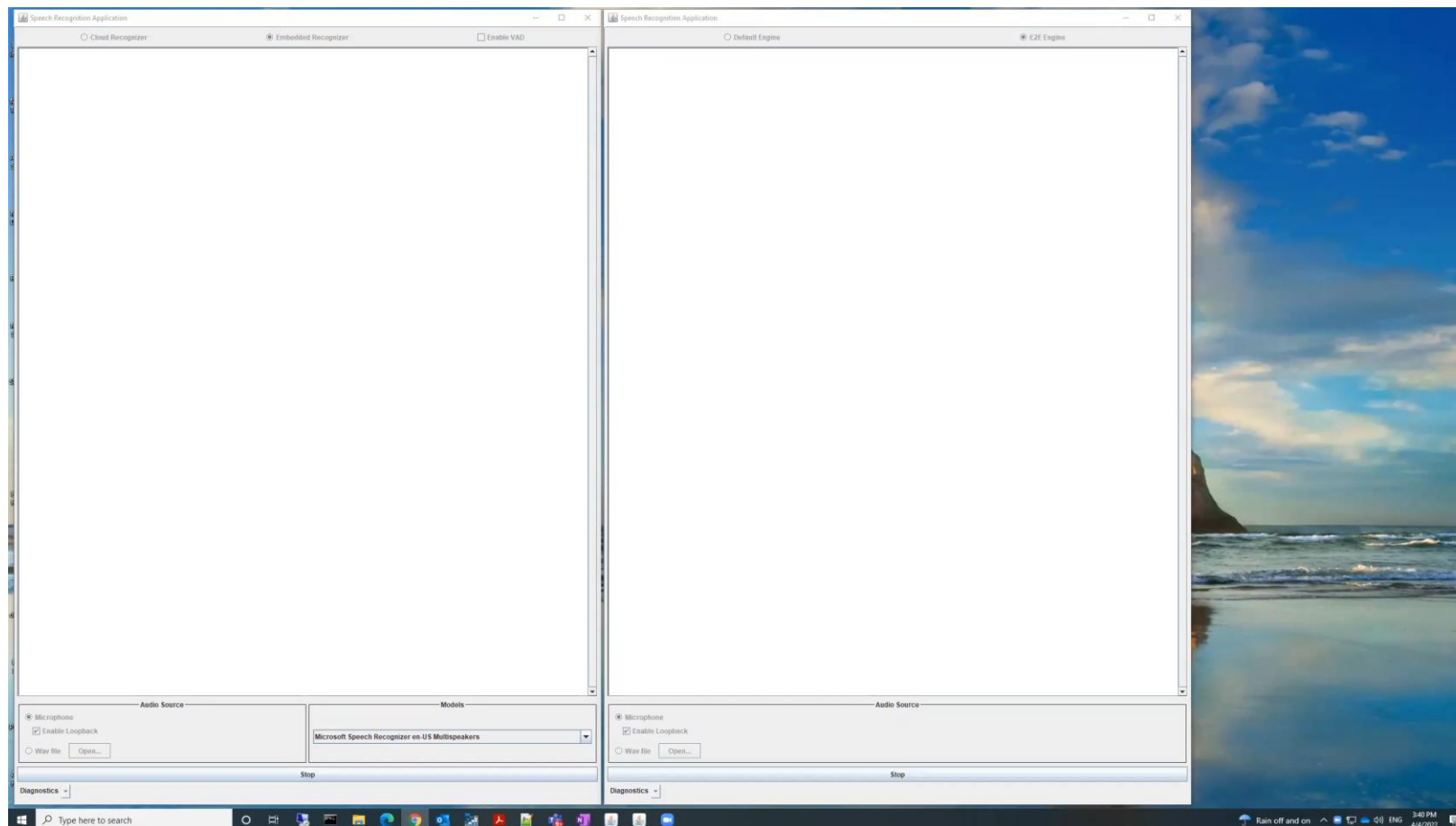
Beyond ASR

E2E Speech Translation (ST)

- ASR is often the first step in a system pipeline and is followed by
  - machine translation
  - speech synthesis (→ speech-to-speech translation)
  - natural language understanding / generation, etc.
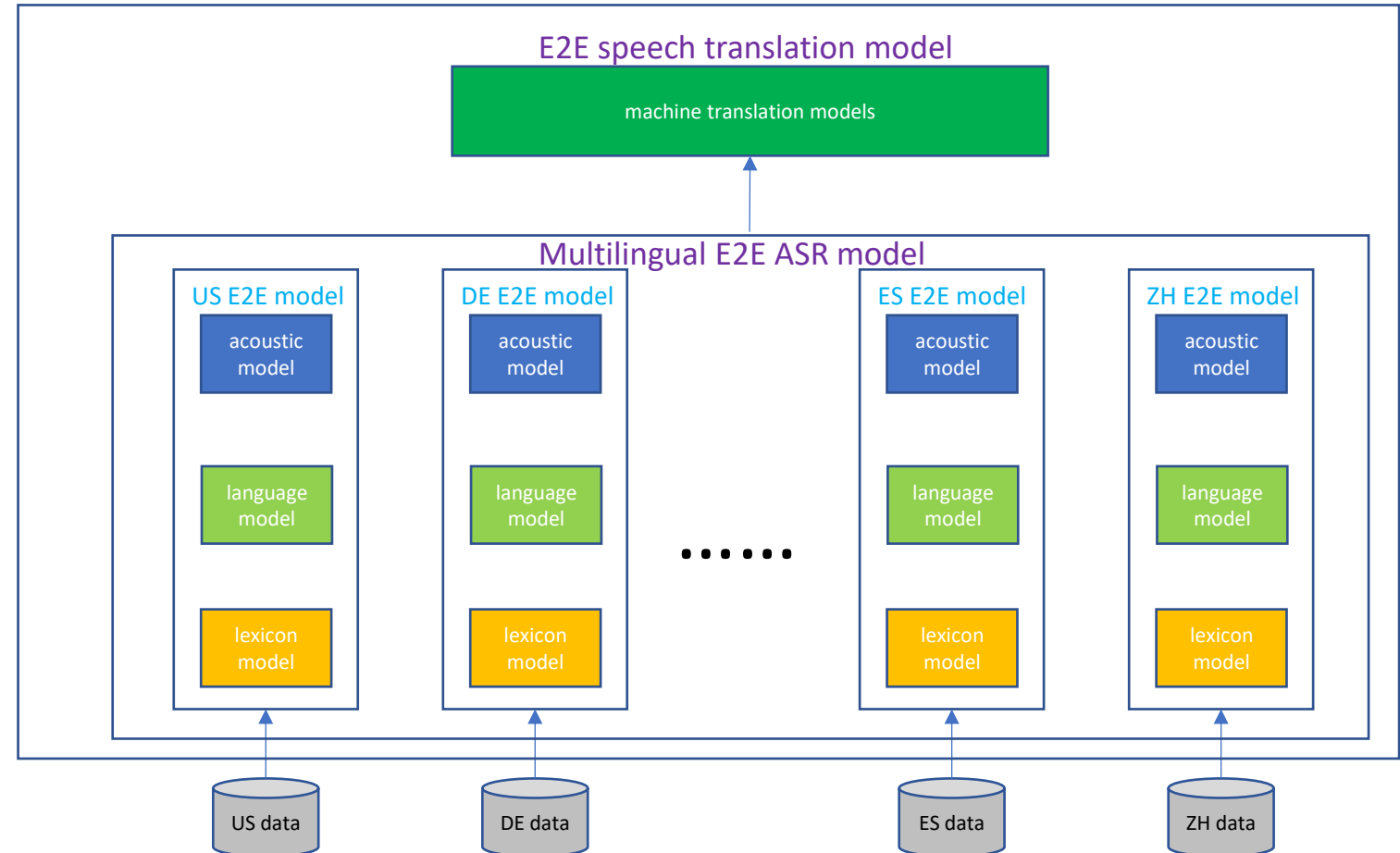
Streaming Multilingual Speech Model (SM^2)
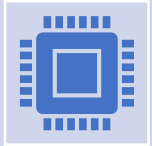
🔲 Multilingual data is pooled together to train a streaming Transformer Transducer model to perform both ST and ASR functions.
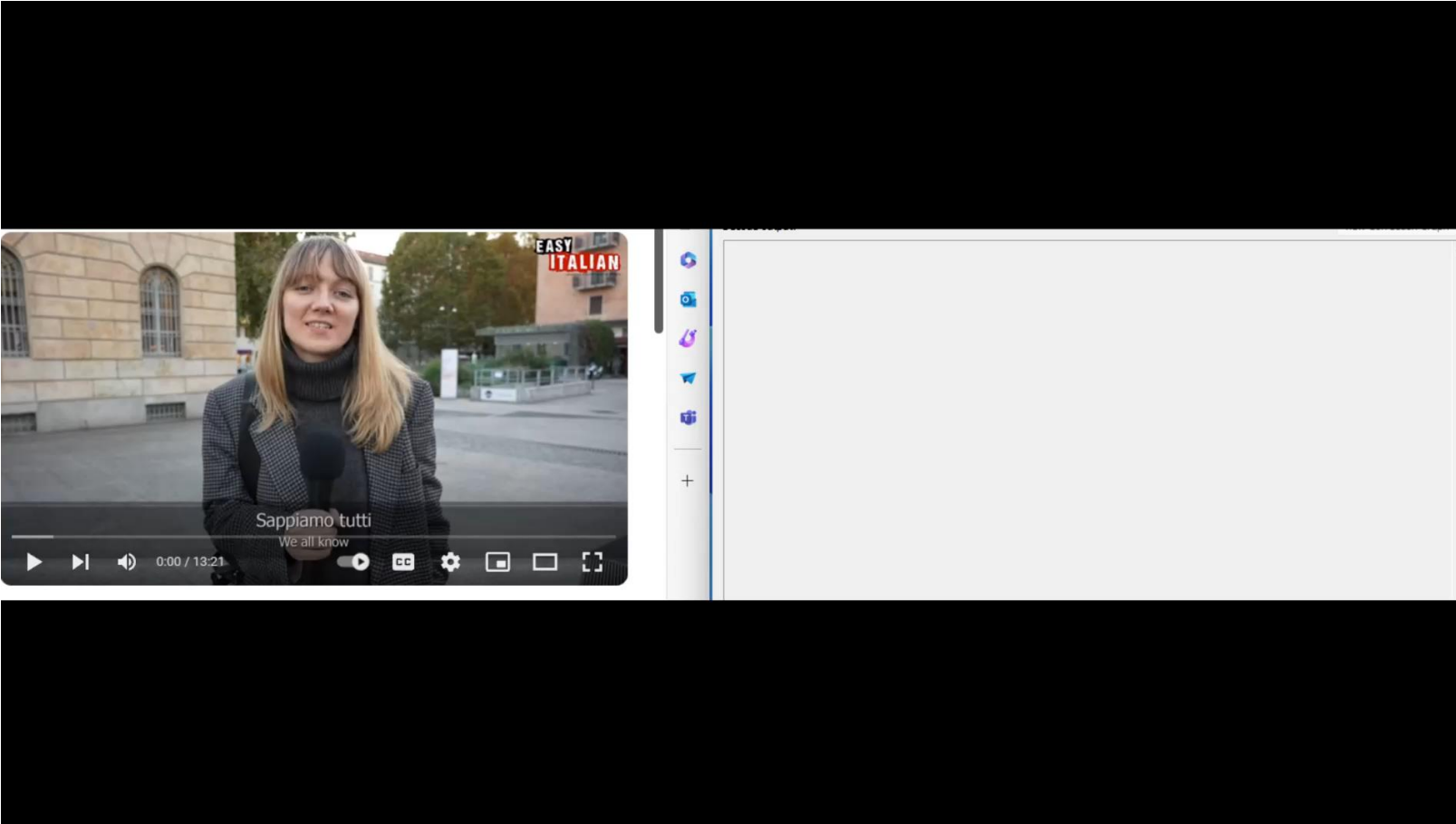
🗄 ST training is totally weakly supervised without using any human labeled parallel corpus.
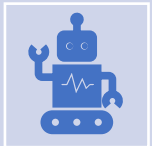
🤖 The model is very small, running on devices with low latency.

J. Xue, et al., "A weakly-supervised streaming multilingual speech model with truly zero-shot capability," In *Proc. ASRU,* 2023.

Simultaneous ST Demo

Available for Public Preview at Microsoft

SDK | CLI | Text to speech service | **Speech to text service** | Containers

April 2024 release

Automatic multi-lingual speech translation (Preview)

Automatic multi-lingual speech translation is available in public preview. This innovative feature revolutionizes the way language barriers are overcome, offering unparalleled capabilities for seamless communication across diverse linguistic landscapes.

Key Highlights

- Unspecified input language: Multi-lingual speech translation can receive audio in a wide range of languages, and there's no need to specify what the expected input language is. It makes it an invaluable feature to understand and collaborate across global contexts without the need for presetting.
- Language switching: Multi-lingual speech translation allows for multiple languages to be spoken during the same session, and have them all translated into the same target language. There's no need to restart a session when the input language changes or any other actions by you.

How it works

- Travel interpreter: multi-lingual speech translation can enhance the experience of tourists visiting foreign destinations by providing them with information and assistance in their preferred language. Hotel concierge services, guided tours, and visitor centers can utilize this technology to cater to diverse linguistic needs.
- International conferences: multi-lingual speech translation can facilitate communication among participants from different regions who might speak various languages using live translated caption. Attendees can speak in their native languages without needing to specify them, ensuring seamless understanding and collaboration.
- Educational meetings: In multi-cultural classrooms or online learning environments, multi-lingual speech translation can support language diversity among students and teachers. It allows for seamless communication and participation without the need to specify each student's or instructor's language.

How to access

For a detailed introduction, visit Speech translation overview. Additionally, you can refer to the code samples at how to translate speech. This new feature is fully supported by all SDK versions from 1.37.0 onwards.

Foundation Model -- Whisper

Trained from 680k hours human caption 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

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

Universal Speech Understanding (USM) model

Microsoft

Better Self-Sup Algorithm: BEST-RQ	A 2B pre-trained encoder, finetuned by your favorite decoder • AED / CTC / RNN-T
Language and Domain expansion	Expanded to 289 languages
Modality Expansion (Speech + Text)	Maestro + Best-RQ
Model Delivery (Adaptation, Modularization)	Joint finetine from speech and text data. USM-CTC: 800x real-time on TPUv4i

Language Expansion (Lower is better)
- SoTA
- Our Best Results
- No Existing Result

Downstream ASR Tasks (Lower is better)
- Whisper (SoTA, w/o in-domain data)
- Our Best Results (w/o in-domain data)
- Our Best Results

Downstream AST Tasks (Higher is better)
- Whisper (SoTA)
- Our Best Results

Slide credit to Yu Zhang

Y. Zhang, 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 AI

Image credit to XD Huang
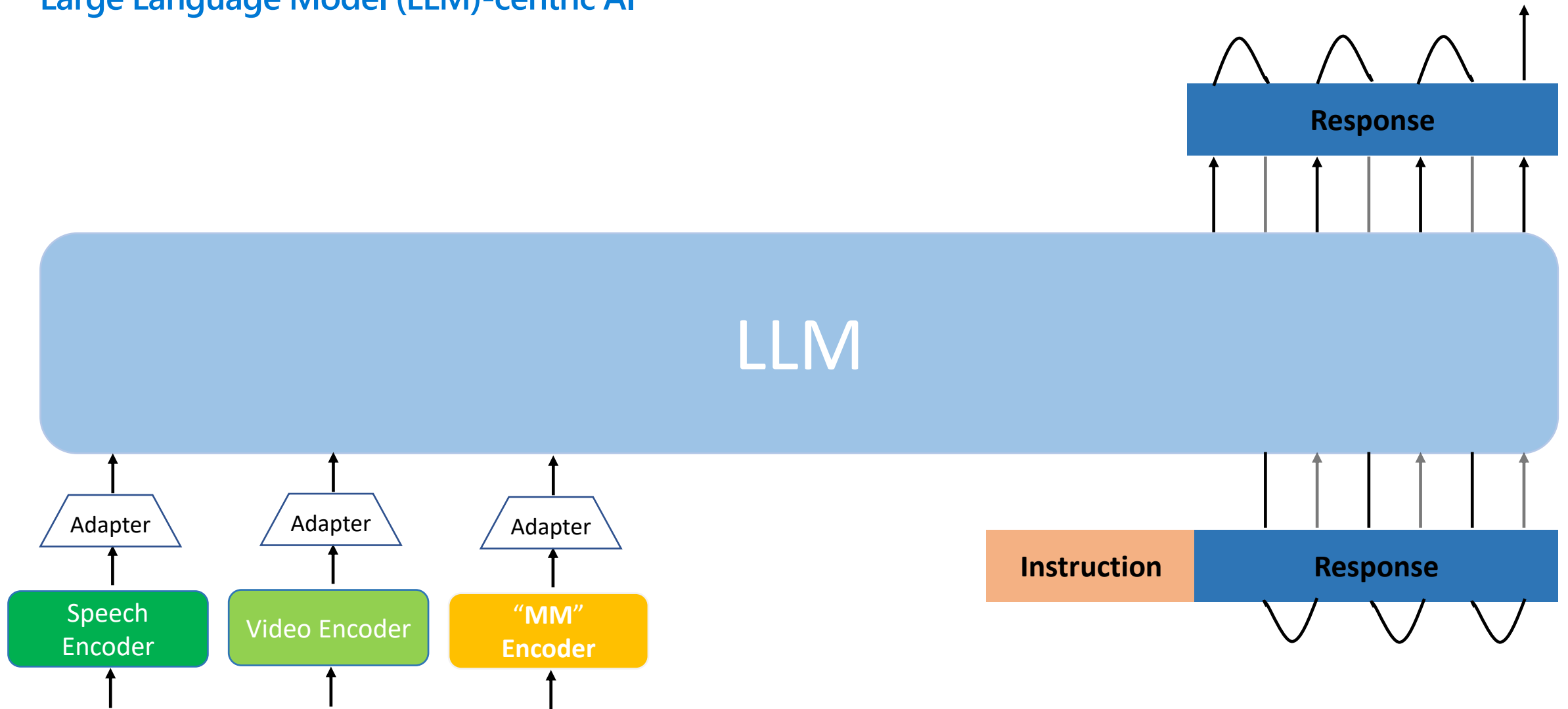
VALL-E: Neural Codec Language Model

- ## High quality zero shot TTS: In context learning through prompts
  - ### "Steal voice from 3 second's prompt"

Model Overview
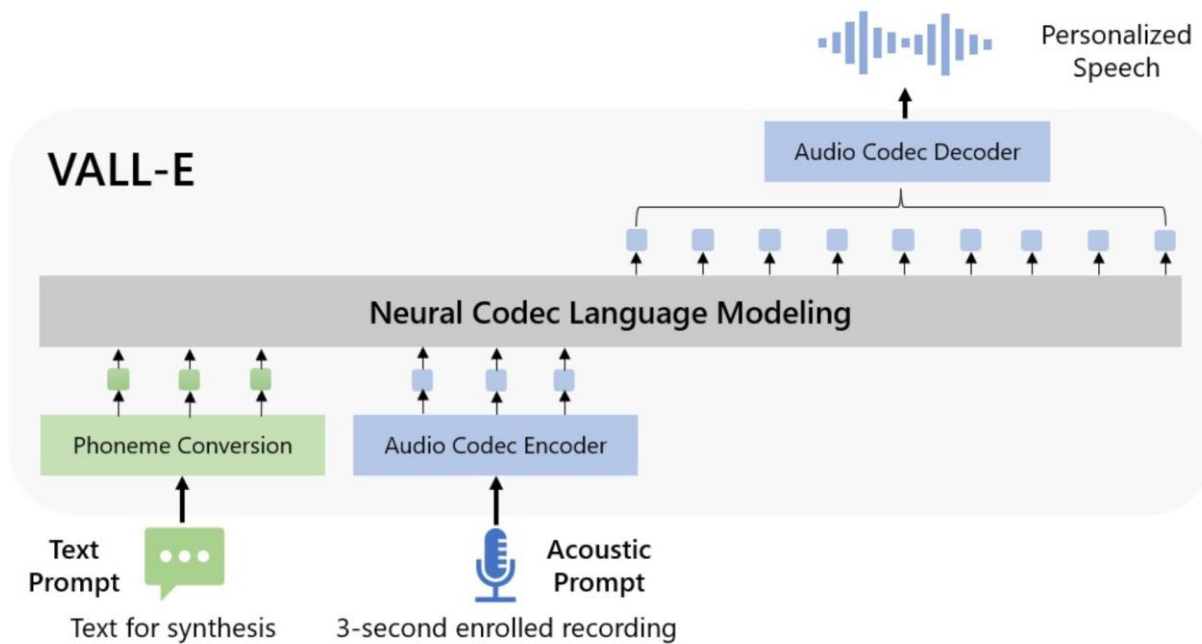
Prompt		Output
🔊		🔊
🔊	I like hamburger but I love noodles much more	🔊
🔊		🔊

C. Wang, 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.
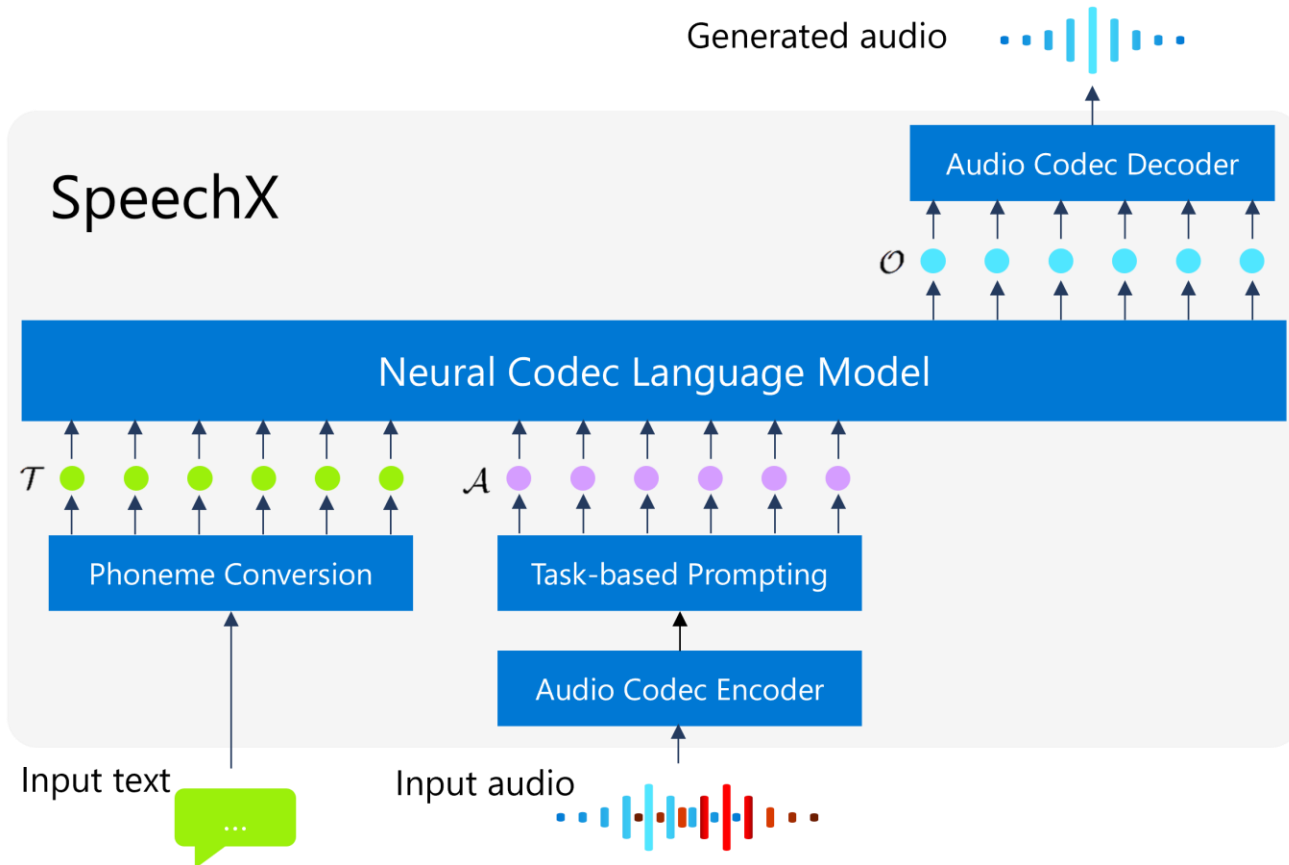
Task	Input text	Input audio	Output audio
Noise suppression	Transcription (optional)	Noisy speech	Clean speech
Speech removal	Transcription (optional)	Noisy speech	Noise
Target speaker extraction	Transcription (optional)	Speech mixture, Enrollment speech	Clean speech of target speaker
Zero-short TTS	Text 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

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

Multi-modal Model with Discrete Audio Inputs: VioLA/AudioPaLM

Target Speech

VIOLA

Audio Codec Decoder

(Source semantic tokens)   (Target semantic tokens)

Multi-Task Codec Language Modeling

Language ID

(Source acoustic tokens)

ASR Task ID

MT Task ID

TTS Task ID   (Target acoustic tokens)

Audio Codec Encoder

Source Speech
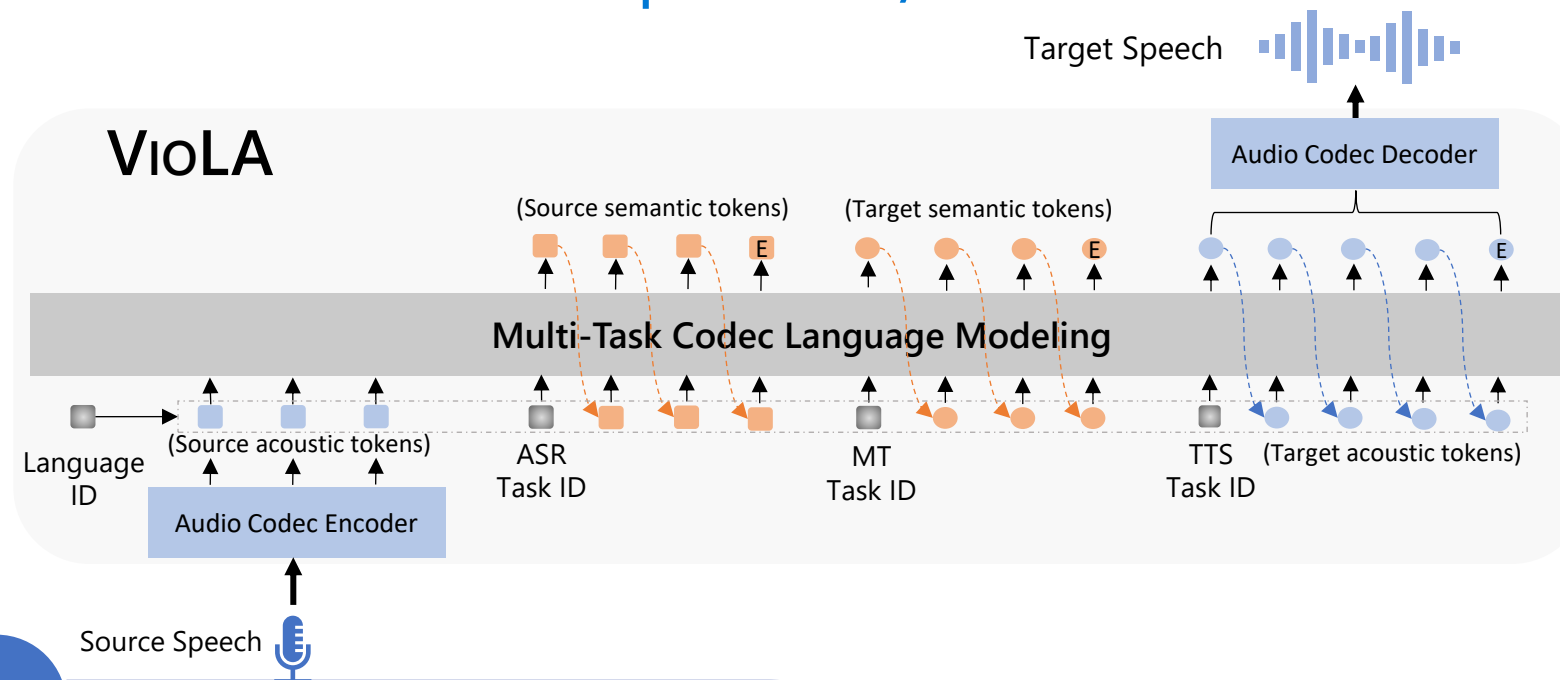
Speech and text can freely serve as input and output

- 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

T. Wang, et al., "VioLA: Unified codec language models for speech recognition, synthesis, and translation," arXiv:2305.16107, 2023.
P.K. Rubenstein, et al., "Audiopalm: A large language model that can speak and listen," arXiv:2306.12925, 2023.

Multi-modal Model with Continuous Audio Inputs: SALMONN/SLM/WavLLM

SALMONN
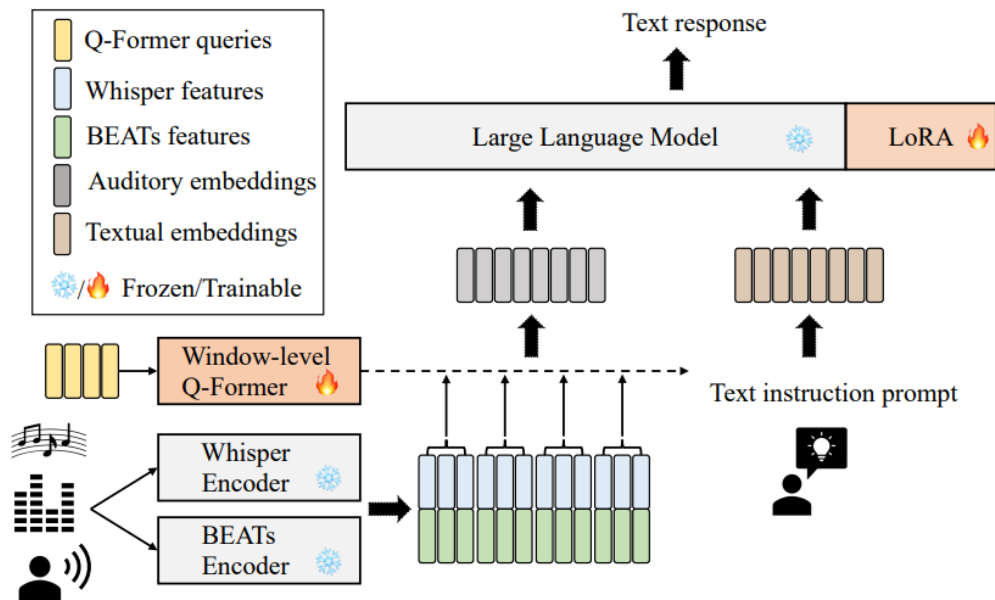
WavLLM
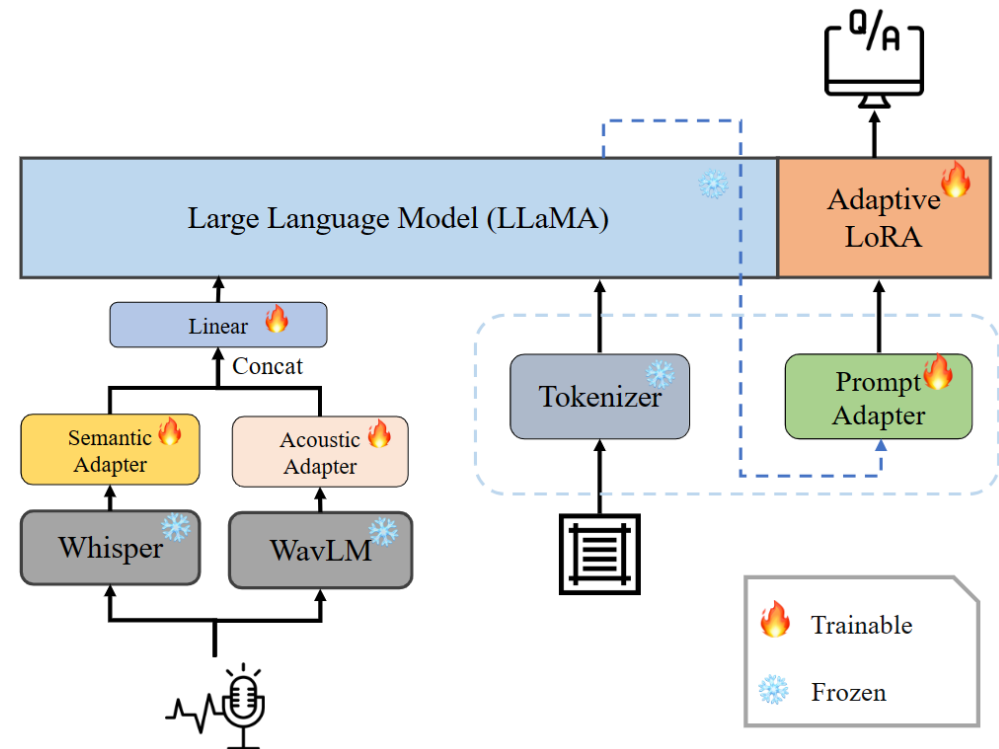
C. Tang, et al., "Salmonn: Towards generic hearing abilities for large language models," arXiv:2310.13289, 2023.

M. Wang, et al., "SLM: Bridge the thin gap between speech and text foundation models, in Proc. ASRU, 2023.

S. Hu, et al., "WavLLM: Towards robust and adaptive speech large language model," *arXiv:2404.00656, 2024*.

Examples of WavLLM

Task	Audio	Prompt	Output
ASR		Based on the attached audio, generate a comprehensive text transcription of the spoken content.	he hoped there would be stew for dinner turnips and carrots and browned potatoes and fat mutton pieces to be ladled out in thick peppered flour fattened sauce
SV		Is there only one speaker in the audio clip?	Incorrect, the audio clip features two individuals speaking.
ST		Translate the audio clip into German.	Sie wird es wohl überleben.
ER		Can you describe the emotional condition of the speaker in the provided audio clip?	sad
SQA		What will the man do next? A. Start to take exercise; B. Do as he always does; C. Change his working time.	A

Examples of WavLLM

Task	Audio	Prompt	Output
SQQA	🔊	N/A	The fundamental theorem of calculus states that different rates of change of a function at a point in the domain of the function are equal. This theorem is used to determine the accumulation of a quantity over a given interval, and to establish relationships between different variables. It is also used to calculate derivatives, which are important in many areas such as optimization, physics, and economics.
COT	🔊	First of all, transcribe the audio recording into text, capturing every spoken word; Additionally given this audio clip and text, can you condense it into a clear, concise summary, no more than 20 words?; Lastly disregarding the sound, translate this English summary into German.	The audio transcription is: three films from Asia pacific are in the running for the coveted golden palms at this year's Cannes Film Festival competing in a field dominated by European productions organizers announced Monday; The audio clip have been summarized as follows: Three films from Asia-Pacific are competing for Golden Palms at this year's Cannes Film Festival, challenging European productions.; The translation of summary is: **Drei Filme aus Asien-Pazifik nehmen an diesem Jahr an den Filmfestspielen von Cannes teil und konkurrieren mit europäischen Produktionen um die Goldenen Palmen.**

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

target text

Speech-LLM

few-shot examples audio   input audio   text Instruction   few-shot text examples

J. Wu, et al., "On decoder-only architecture for speech-to-text and large language model integration," *In Proc. ASRU*, 2023.
J. Pan, et al. "COSMIC: Data efficient instruction-tuning for speech in-context learning," *arXiv preprint*, 2023.
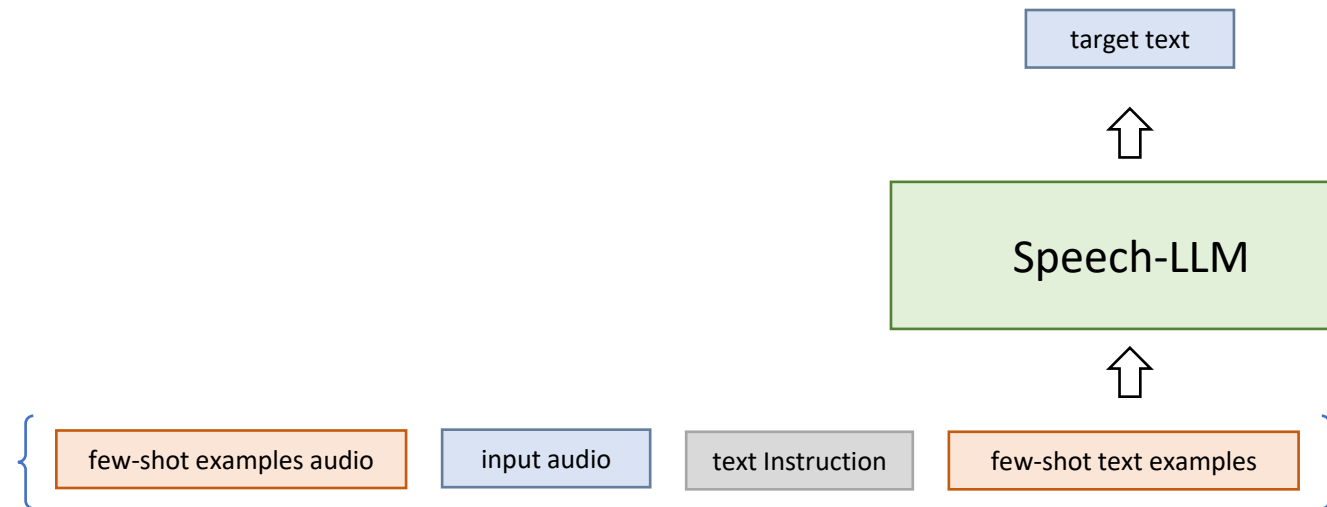
Conclusions

- E2E models are now the mainstream 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
  - Multi-talker ASR: (token-level) serialized output training
  - Beyond ASR: streaming multilingual speech model

- Large language model (LLM) centric AI may be the next trend.

Reference

- Most of contents are based on

J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing* 11, 2022.

Industrial Technology/Advances

Recent Advances in End-to-End Automatic Speech Recognition

Jinyu Li*

Microsoft, Redmond, WA 98052, USA

ABSTRACT

Recently, the speech community is seeing a significant trend of moving from deep neural network based hybrid modeling to end-to-end (E2E) modeling for automatic speech recognition (ASR). While E2E models achieve the state-of-the-art results in most benchmarks in terms of ASR accuracy, hybrid models are still used in a large proportion of commercial ASR systems at the current time. There are lots of practical factors that affect the production model deployment decision. Traditional hybrid models, being optimized for production for decades, are usually good at these factors. Without providing excellent solutions to all these factors, it is hard for E2E models to be widely commercialized. In this paper, we will overview the recent advances in E2E models, focusing on technologies addressing those challenges from the industry's perspective.

Reference

- N. Arivazhagan et al., "Monotonic infinite lookback attention for simultaneous machine translation," in Proc. ACL, 2019.

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

- X. Chen et al., "Factorized neural transducer for efficient language model adaptation," in Proc. ICASSP, 2022.

- C. Chiu and C. Raffel, "Monotonic chunkwise attention," in Proc. ICLR, 2018.

- J. Chorowski, et al., "End-to-end continuous speech recognition using attention-based recurrent NN: First results," arXiv:1412.1602, 2014.

- A. Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proc. ICML, 2006.

- A. Graves., "Sequence transduction with recurrent neural networks," in arXiv preprint, 2012.

- A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks" PMLR, 2014.

- M. Ghodsi et al., "RNN-Transducer with stateless prediction network," in Proc. ICASSP, 2020.

- A. Gulati, et al., "Conformer: Convolution-augmented Transformer for speech recognition," in Proc. Interspeech, 2020.

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

- A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," in arXiv preprint, 2014.

- S. Hu, et al., "WavLLM: Towards robust and adaptive speech large language model," arXiv:2404.00656, 2024.

- N. Kanda, et al., "Serialized output training for end-to-end overlapped speech recognition," In Proc. Interspeech, 2020.
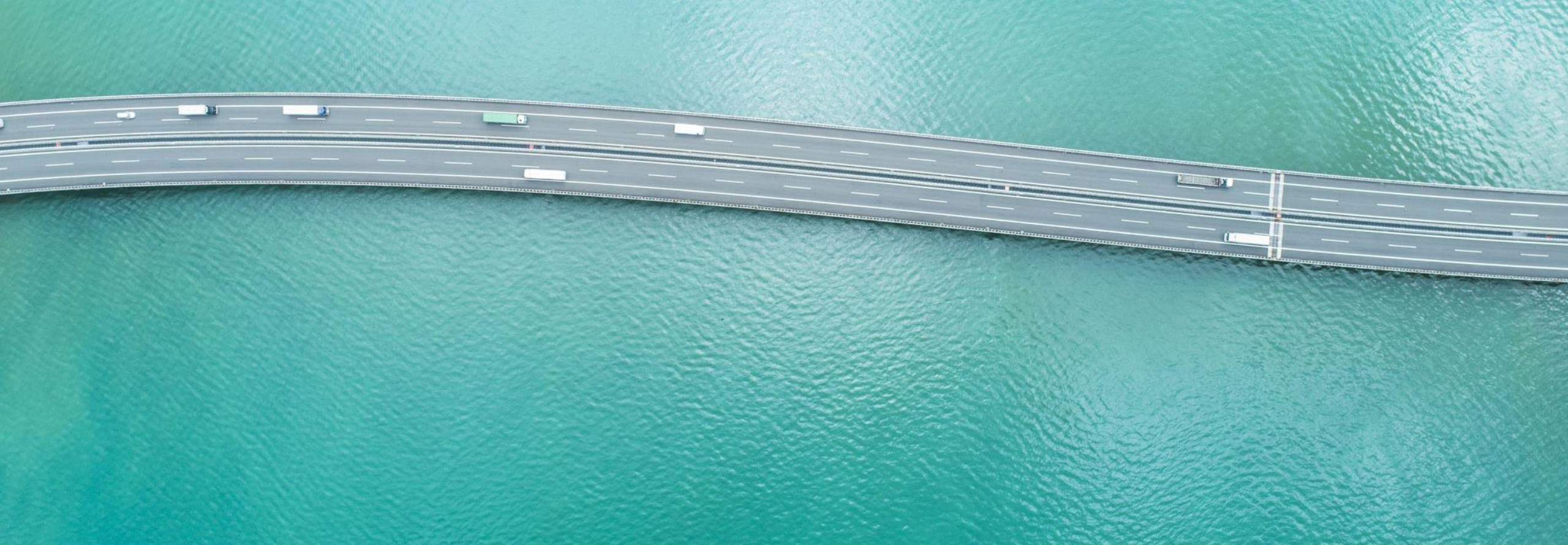
Reference

- N. Kanda, et al., "Streaming speaker-attributed ASR with token-level speaker embeddings." in Proc. Interspeech, 2022.

- S. Kim and M. Seltzer, "Towards language-universal end-to-end speech recognition," in Proc. ICASSP, 2018.

- J. Li, et al., "Developing RNN-T models surpassing high-performance hybrid models with customization capability," in Proc. Interspeech, 2020.

- J. Li, "Recent advances in end-to-end automatic speech recognition," APSIPA Transactions on Signal and Information Processing 11, 2022.

- L. Lu, et al., "Streaming end-to-end multi-talker speech recognition," IEEE Signal Processing Letters, 2021.

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

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

- J. Pan, et al. "COSMIC: Data efficient instruction-tuning for speech in-context learning," arXiv preprint, 2023.

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

- P.K. Rubenstein, et al., "Audiopalm: A large language model that can speak and listen," arXiv:2306.12925, 2023.

- T. Sainath, et al., "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in Proc. ICASSP, 2020.

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

- I. Sklyar, et al., "Streaming multi-speaker ASR with RNN-T," in Proc. ICASSP, 2021.

- C. Tang, et al., "Salmonn: Towards generic hearing abilities for large language models," arXiv:2310.13289, 2023.

Reference

- S. Toshniwal et al., "Multilingual speech recognition with a single end-to-end model," in Proc. ICASSP, 2018.

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

- A. Vaswani, et al., "Attention is all you need," NIPS, 2017.

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

- M. Wang, et al., "SLM: Bridge the thin gap between speech and text foundation models, in Proc. ASRU, 2023.

- T. Wang, et al., "VioLA: Unified codec language models for speech recognition, synthesis, and translation," arXiv:2305.16107, 2023.

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

- S. Watanabe et al., "Language independent end-to-end architecture for joint language identification and speech recognition," in Proc. ASRU, 2017.

- J. Wu, et al., "On decoder-only architecture for speech-to-text and large language model integration," In Proc. ASRU, 2023.

- J. Xue, et al., "A weakly-supervised streaming multilingual speech model with truly zero-shot capability," In Proc. ASRU, 2023.

- Q. Zhang, et al., "Transformer transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss," in Proc. ICASSP, 2020.

- Y. Zhang, et al., "Google USM: scaling automatic speech recognition beyond 100 languages," arXiv:2303.01037, 2023.

- R. Zhao, et al., "On addressing practical challenges for RNN-Transducer," in Proc. ASRU, 2021.

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

- L. Zhou, et al., "A configurable multilingual model is all you need to recognize all languages," in Proc. ICASSP, 2022

Thank You!