



Advancing End-to-End Automatic Speech Recognition and Beyond

Jinyu Li

Outline

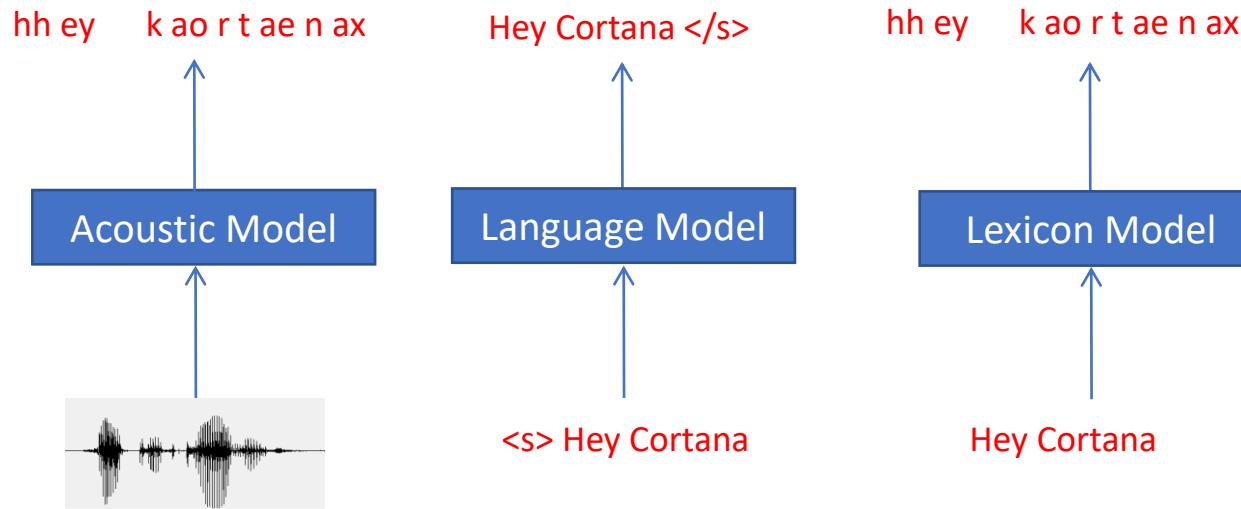
- End-to-end (E2E) automatic speech recognition (ASR) fundamental
- E2E advances
 - Multilingual ASR
 - Leveraging unpaired text
 - Multi-talker ASR
 - Speech translation
- Conclusions and future directions

End-to-End Fundamental

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

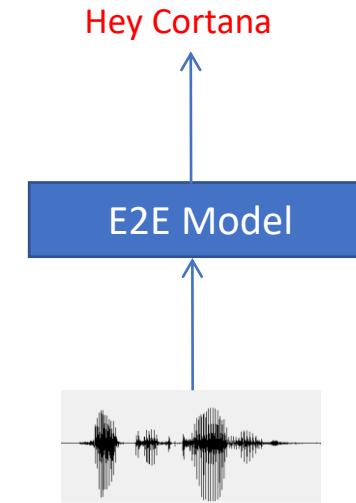
Hybrid

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.



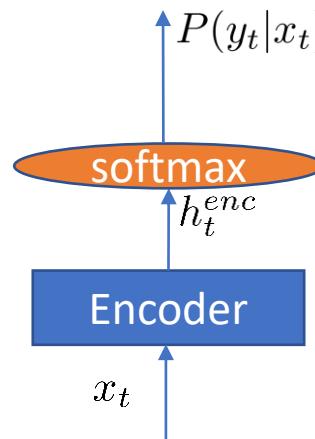
Advantages of E2E Models

- 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

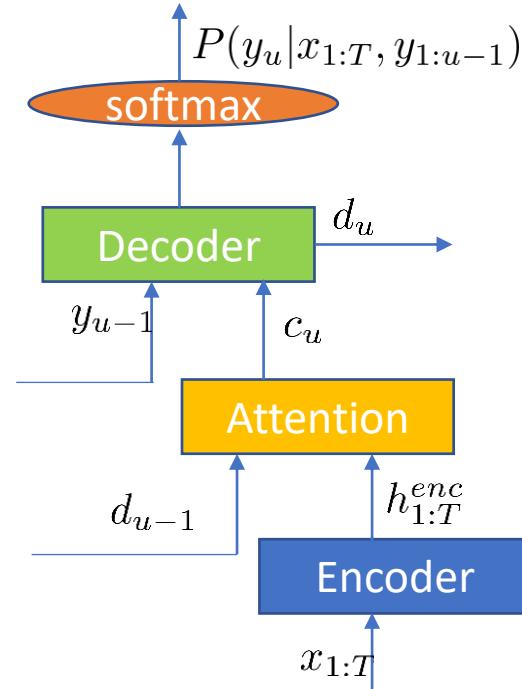
Current Status

- 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.

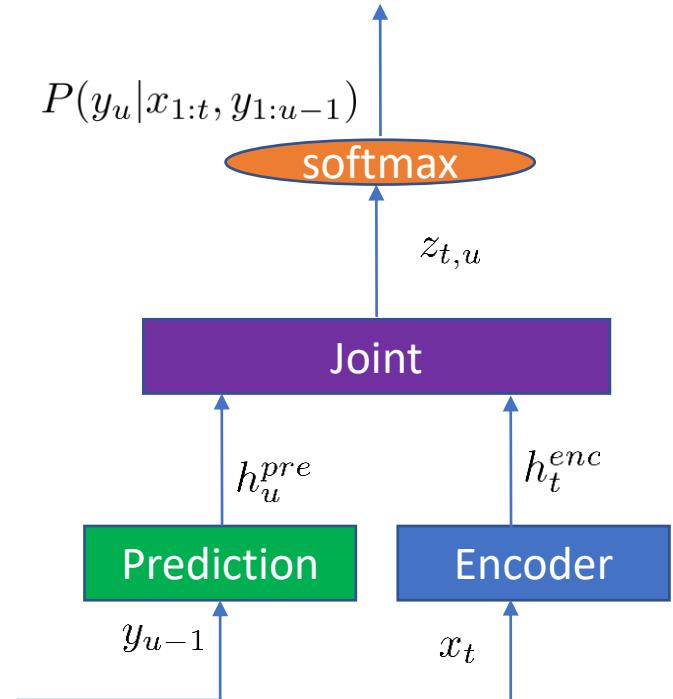
E2E Models



Connectionist Temporal
Classification (CTC)



Attention-based encoder
decoder (AED)



RNN-Transducer
(RNN-T)

CTC

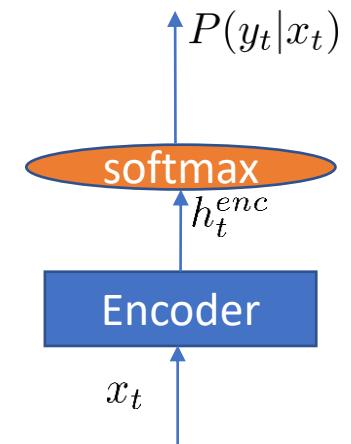
- The first and simplest E2E ASR model.
- To solve the challenge that target label length is smaller than the speech input length:
 - Inserts blank and allows label repetition to have the same length of CTC path and speech input sequence.

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

- 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

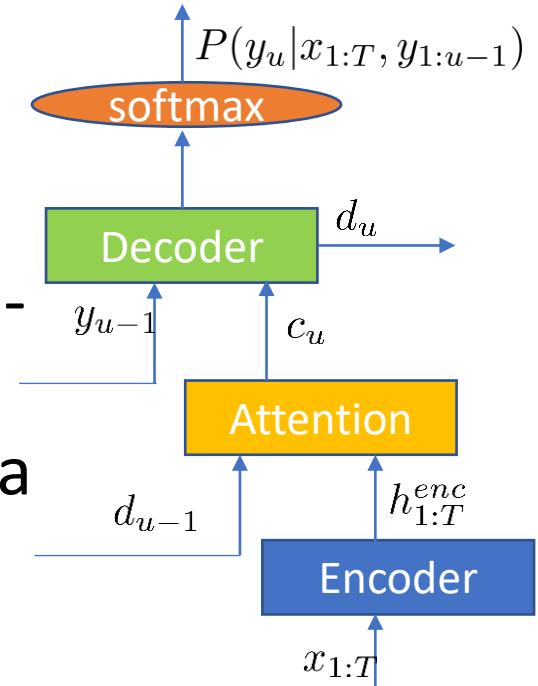


AED

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



Streaming

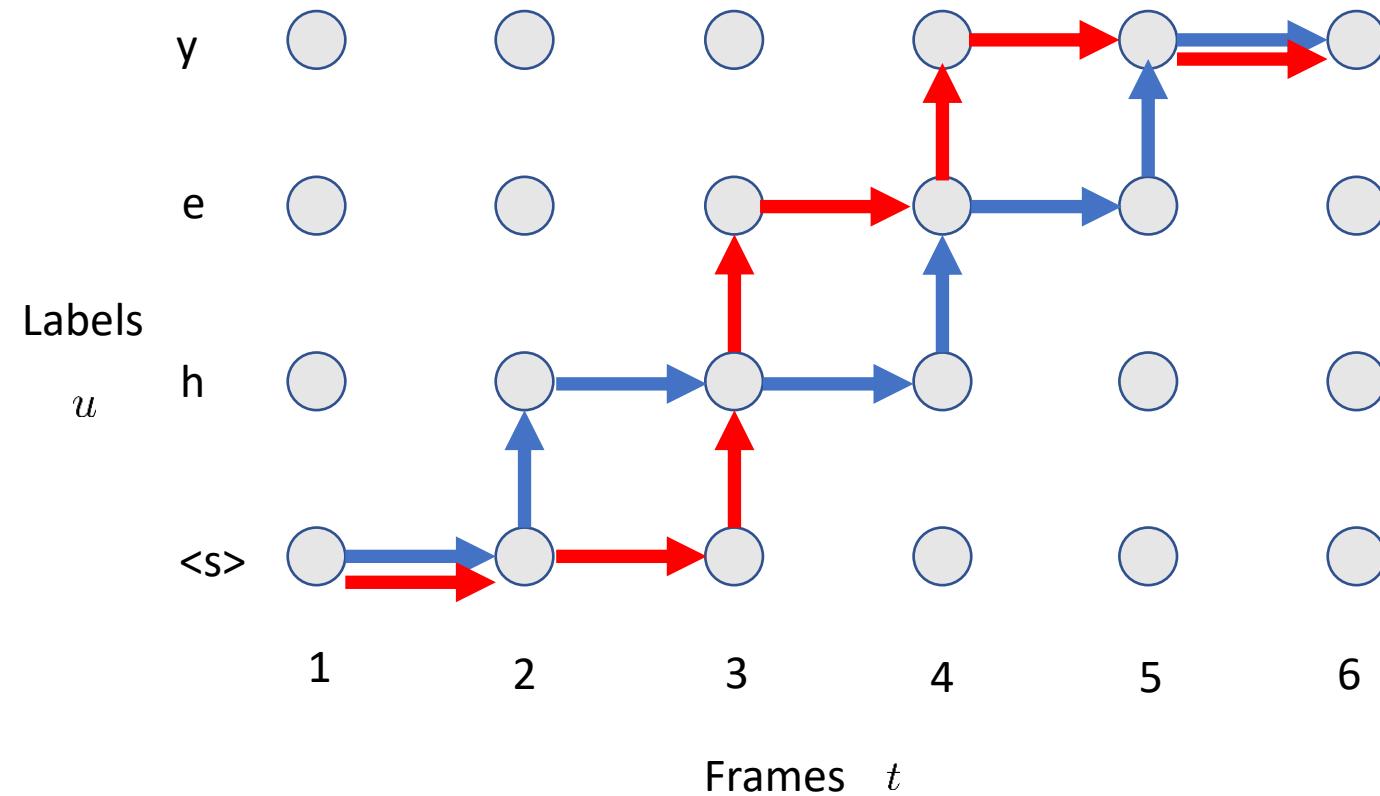
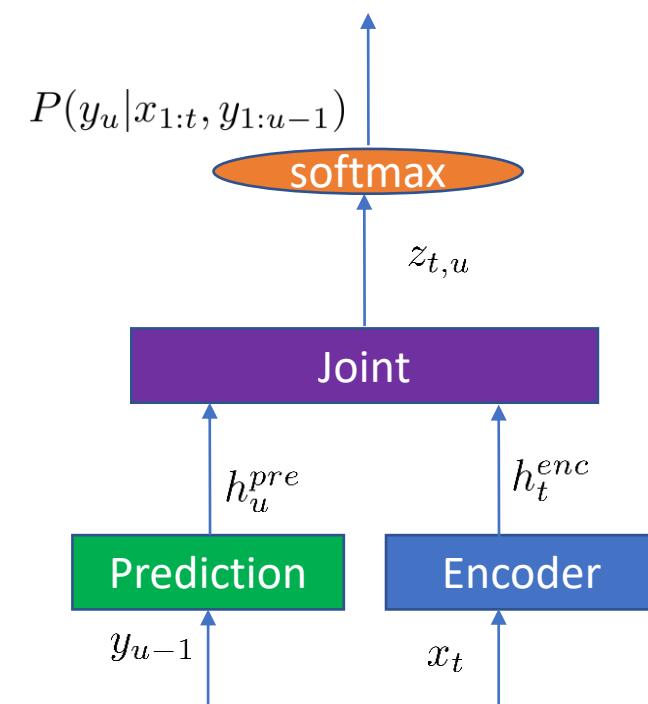
- Most commercial setups need the ASR systems to be streaming with low latency: ASR system produces the recognition results at the same time as the user is speaking.
- Full attention in AED may not be ideal to ASR because the speech signal and output label sequence are monotonic.
 - Streaming AED (MOCHA, 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.

Chiu and Raffel, “Monotonic chunkwise attention,” in Proc. ICLR, 2018.

Arivazhagan et al., “Monotonic infinite lookback attention for simultaneous machine translation,” in Proc. ACL, 2019.

RNN-T

Given a label sequence of length U and acoustic frames T , we generate $U \times T$ softmax. The training maximizes the probabilities of all RNN-T paths.

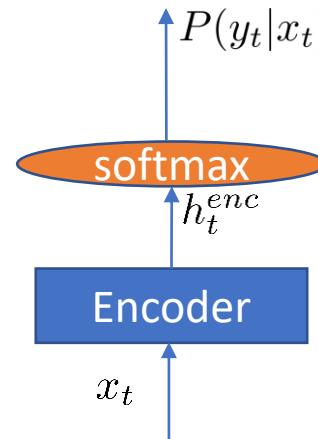


E2E Models

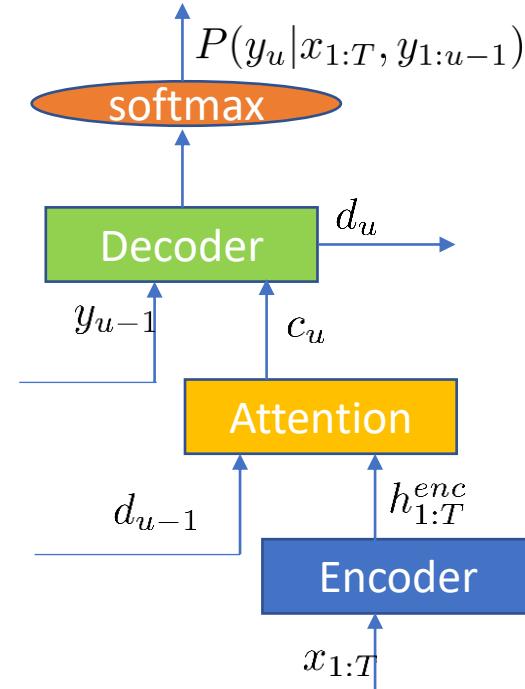
	CTC	AED	RNN-T
Independence assumption	Yes	No	No
Attention mechanism	No	Yes	No
Streaming	Natural	Additional work needed	Natural
Ideal operation scenario	Streaming	Offline	Streaming
Long form capability	Good	Weak	Good

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

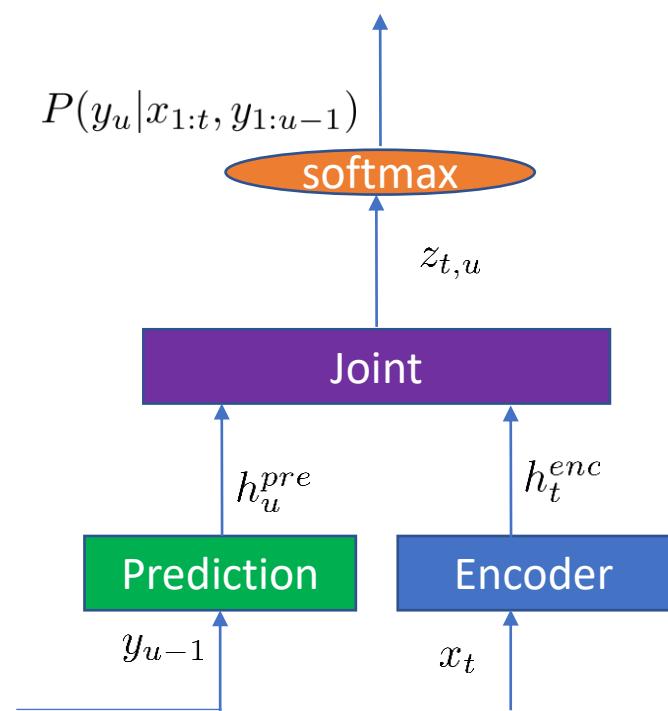
Encoder is the Most Important Component



Connectionist Temporal
Classification (CTC)

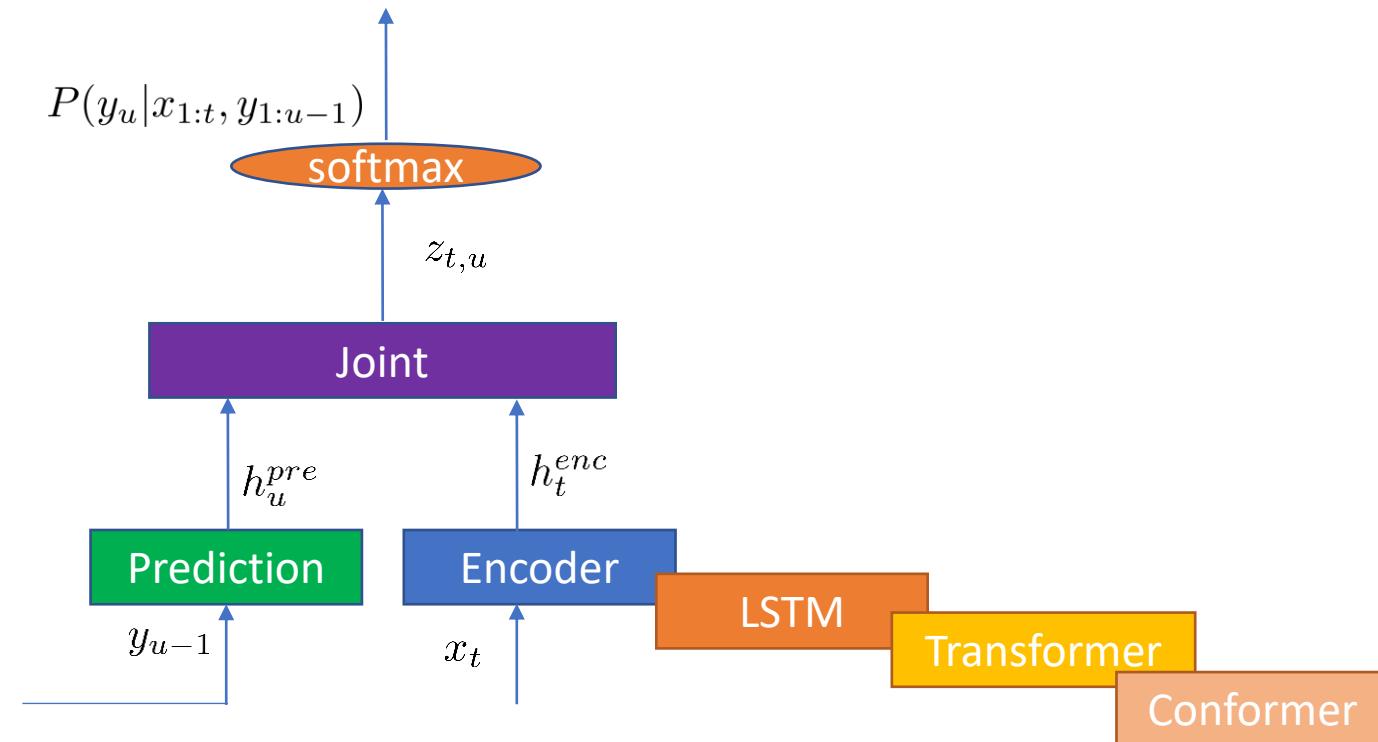


Attention-based encoder
decoder (AED)



RNN-Transducer
(RNN-T)

Encoder for RNN-T



Transformer

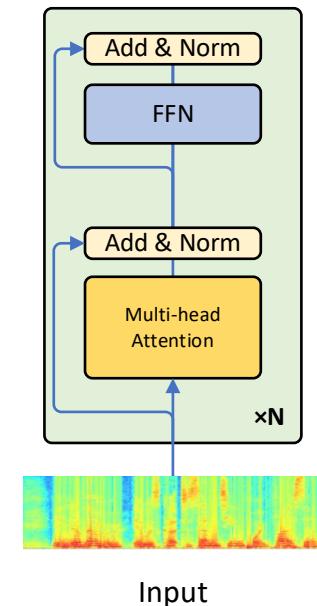
- Self-attention: computes the attention distribution over the input speech sequence

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

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

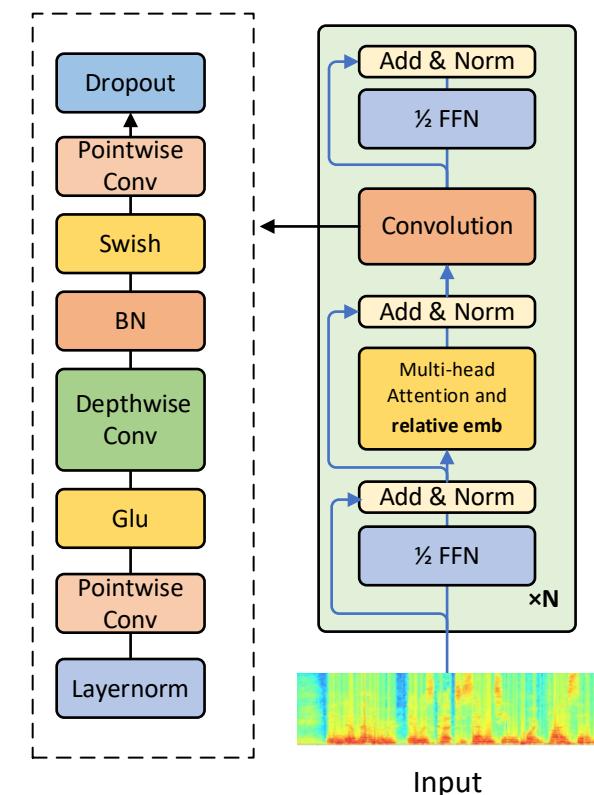
$$\mathbf{z}_t = \sum_{\tau} \alpha_{t\tau} \mathbf{W}_v \mathbf{x}_\tau = \sum_{\tau} \alpha_{t\tau} \mathbf{v}_\tau$$

- 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

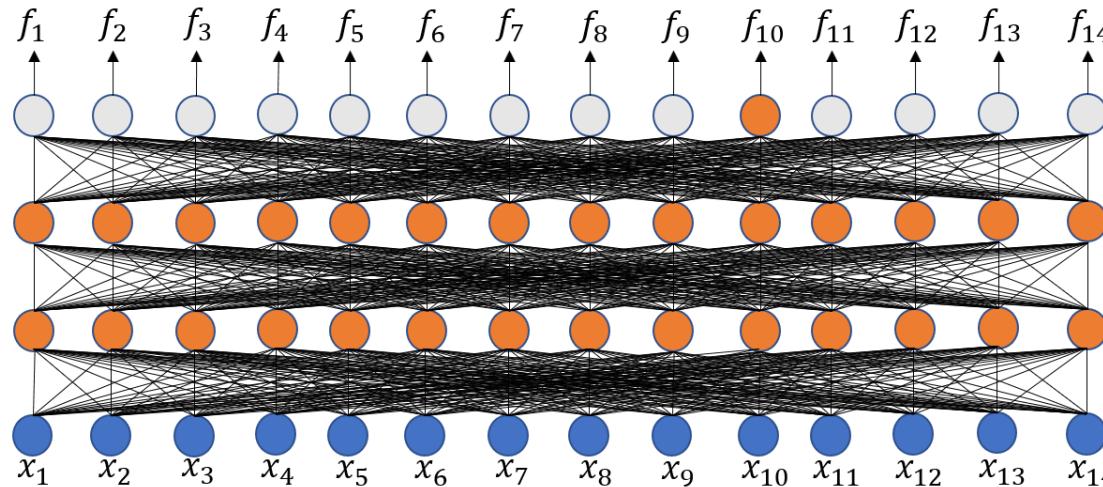
$$\alpha_{t,\tau} = \frac{m_{t,\tau} \exp(\beta (W_q x_t)^T (W_k x_\tau))}{\sum_{\tau'} m_{t,\tau'} \exp(\beta (W_q x_t)^T (W_k x_{\tau'}))} = \text{softmax}(\beta q_t^T k_\tau, m_{t,\tau})$$

- 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}

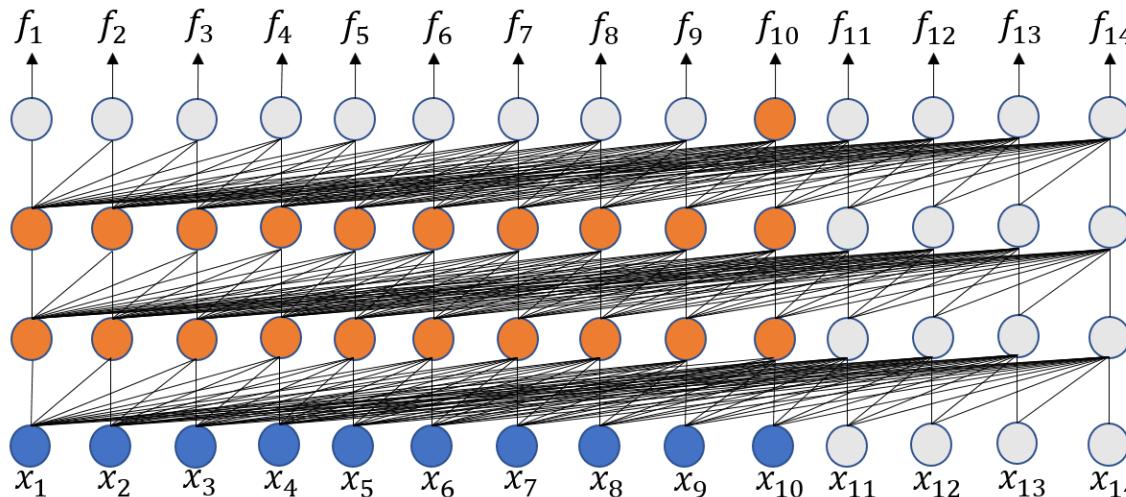
Not streamable

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	1	1
2	1	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	1
4	1	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	1
6	1	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	1
8	1	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	1
10	1	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	1
12	1	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	1
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, full history



Predicting output for x_{10}

**Memory and runtime cost
increase linearly**

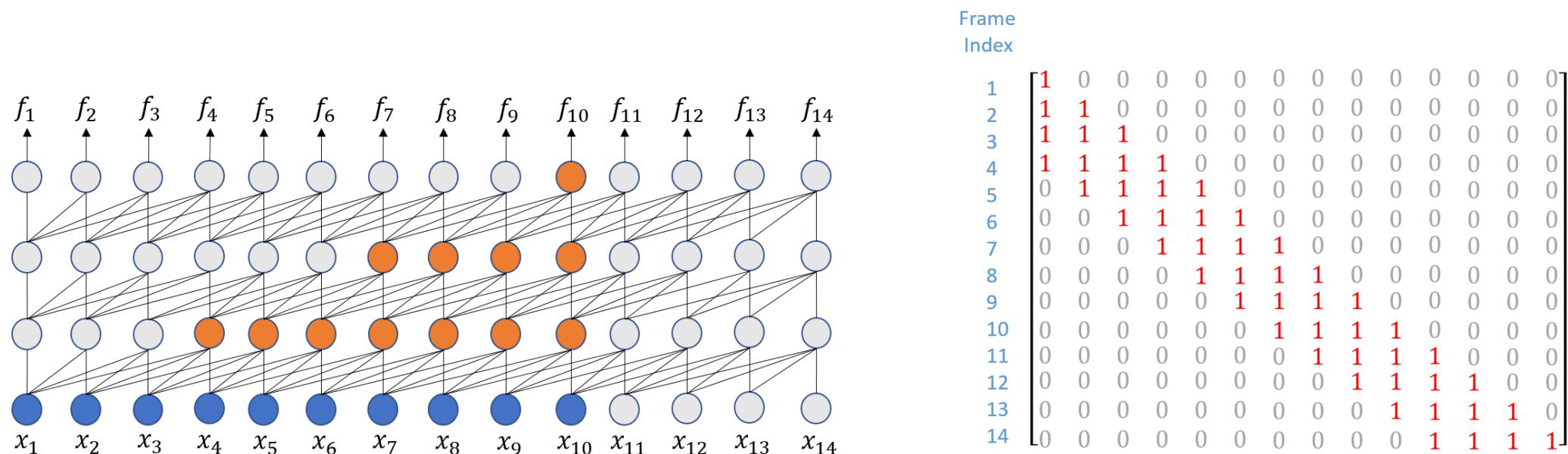
Frame Index

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)



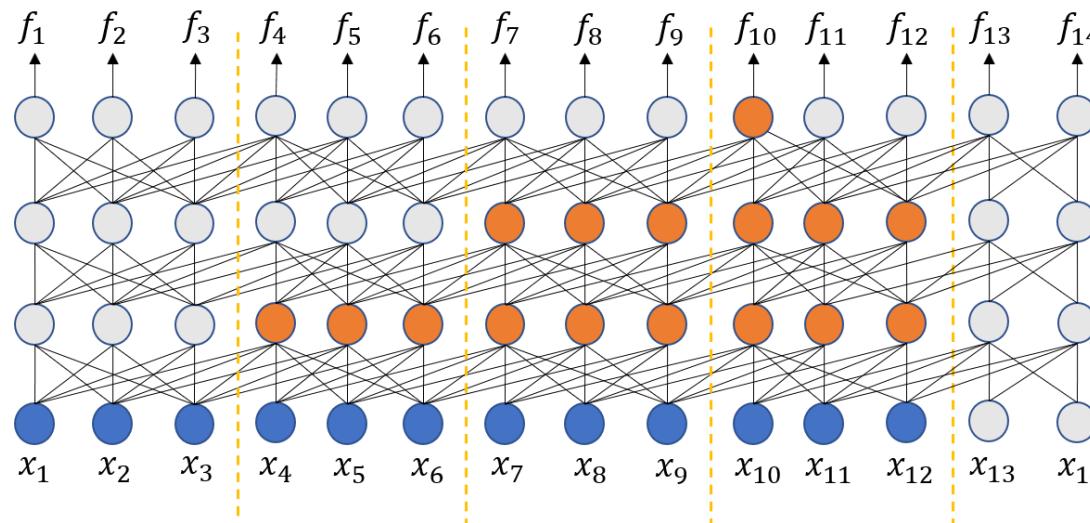
Predicting output for x_{10}

**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)

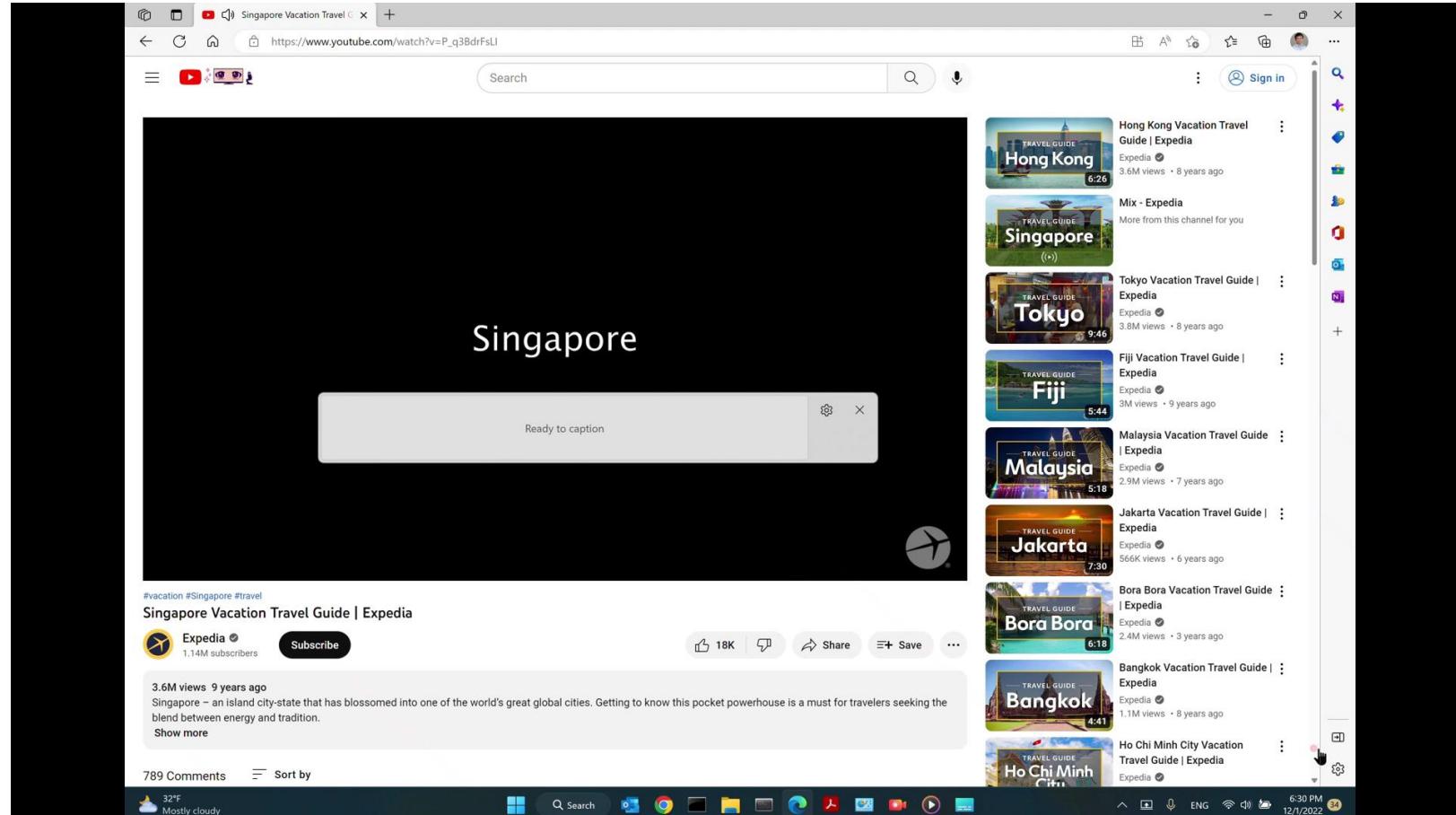


Predicting output for x_{10}

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

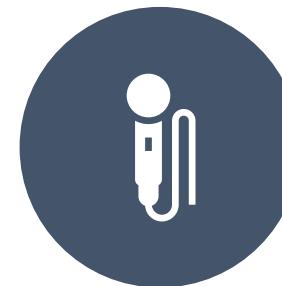
Live Caption in Windows 11



Advancing E2E Models



multilingual



multi-talker



unpaired text



speech translation



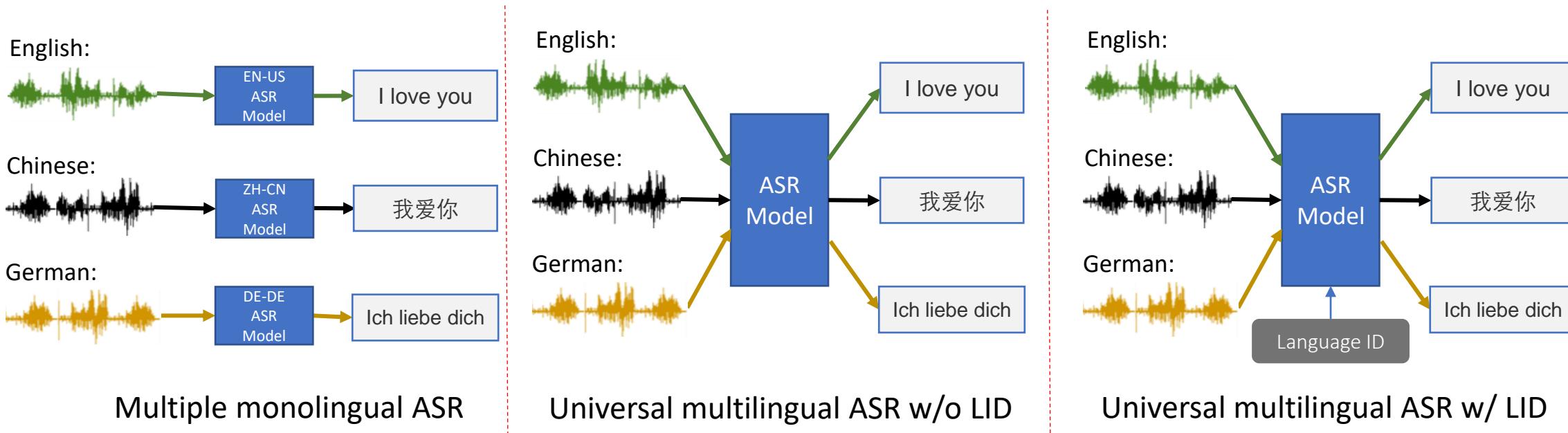
Multilingual

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 on device* 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
 - can be addressed with a one-hot LID input. However, it is more like a monolingual model with the requirement of prior knowledge of language to speak.



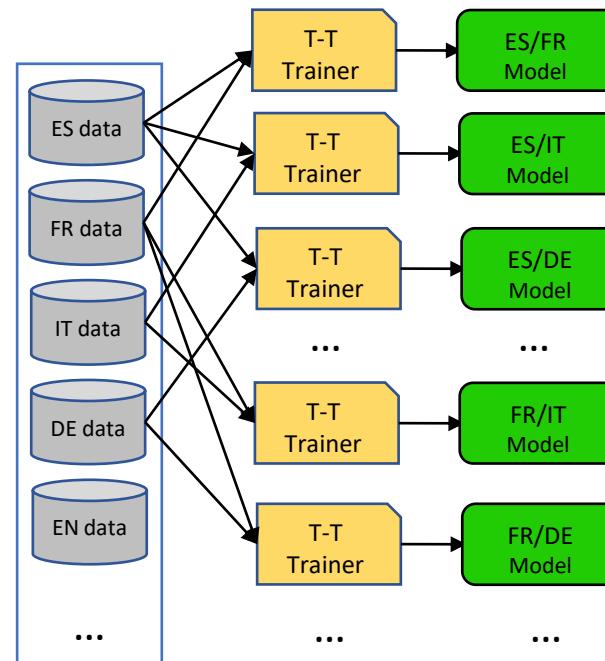
Watanabe et al., "Language independent end-to-end architecture for joint language identification and speech recognition," in Proc. ASRU, 2017.

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

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

Specific Model for Every Combination of Languages?

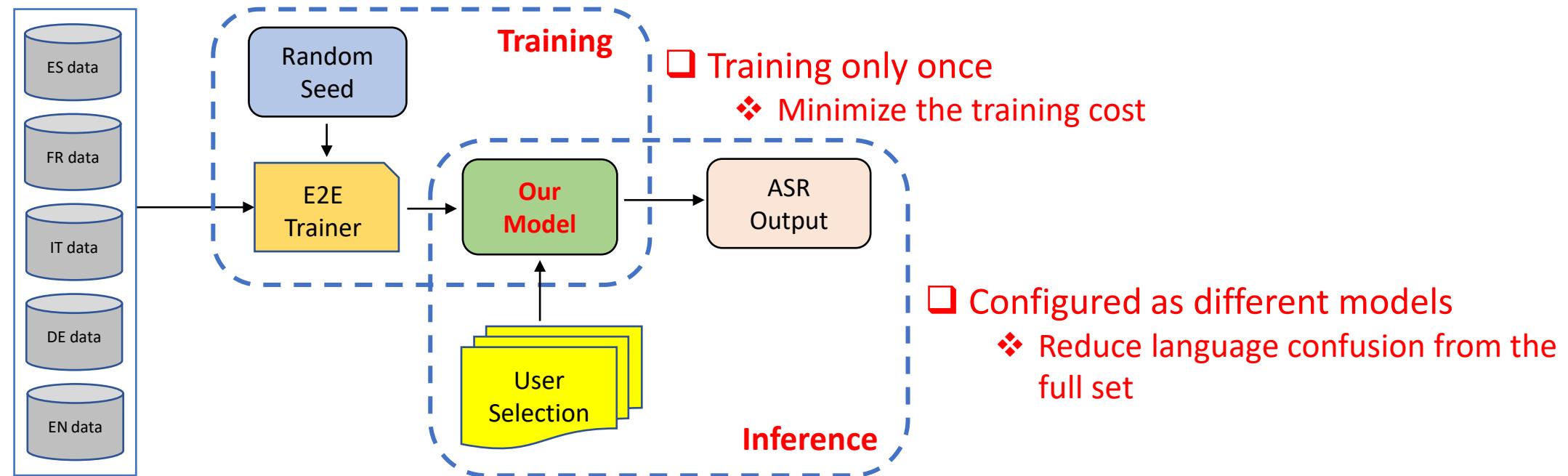
- Advantage: Don't have the confusion from other languages not in target group
- Disadvantage: Development cost is formidable



$\sum_{1 \leq m \leq n} C_n^m$ for n languages
 $C_{10}^1=10, C_{10}^2=45, C_{10}^3=120$

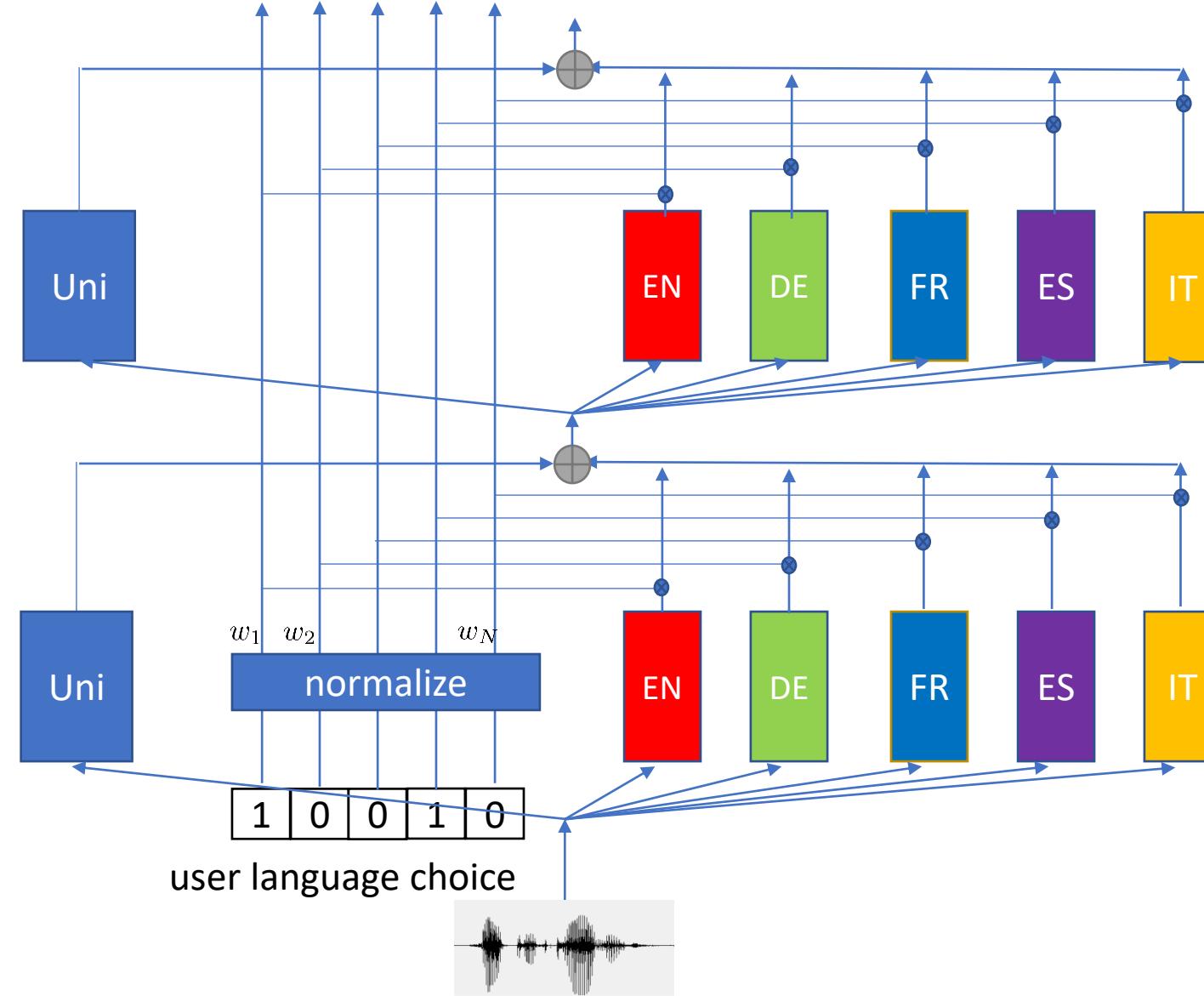
How to Deal with Multilingual Speakers?

Configurable Multilingual Model (CMM)



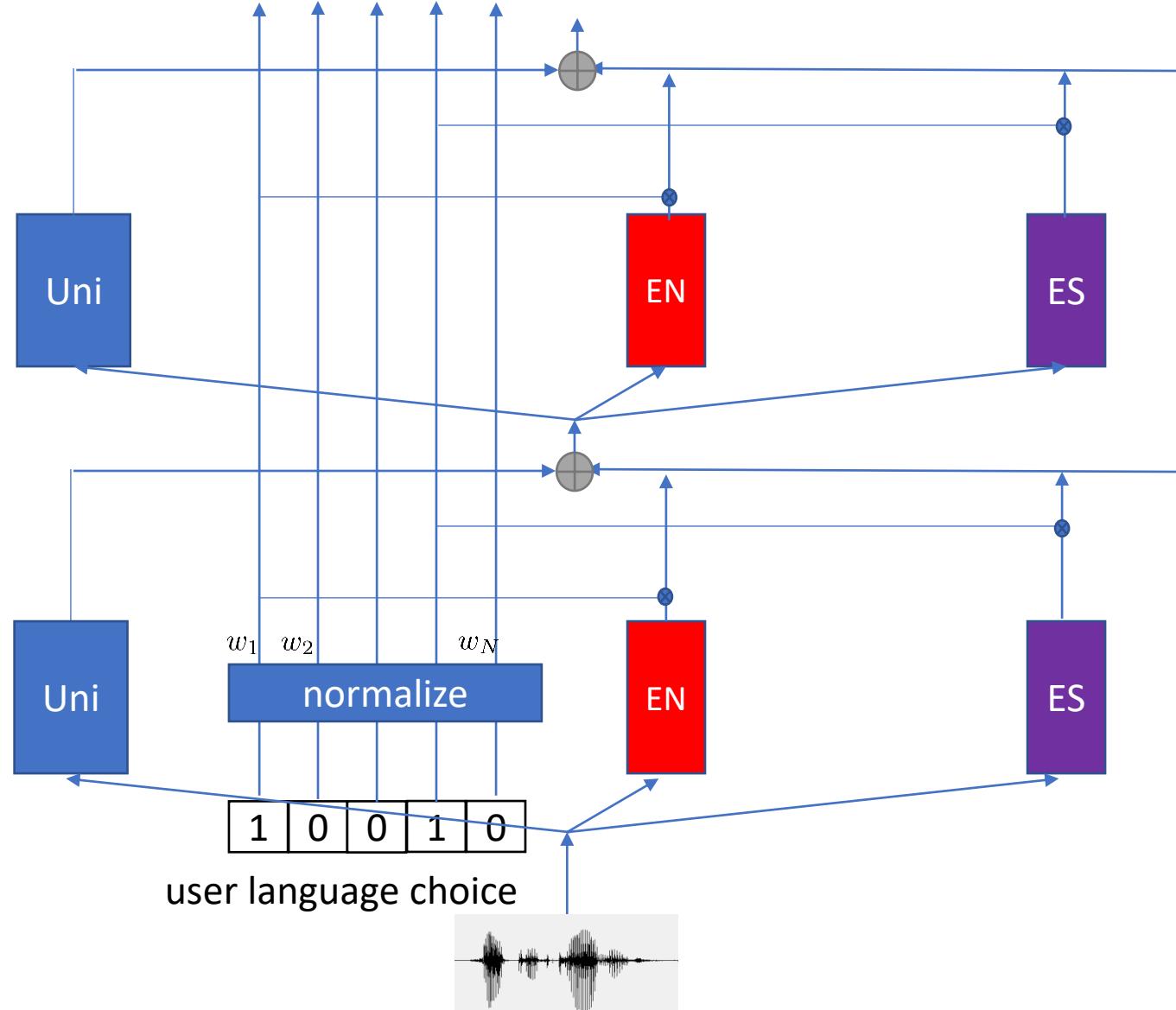
CMM

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



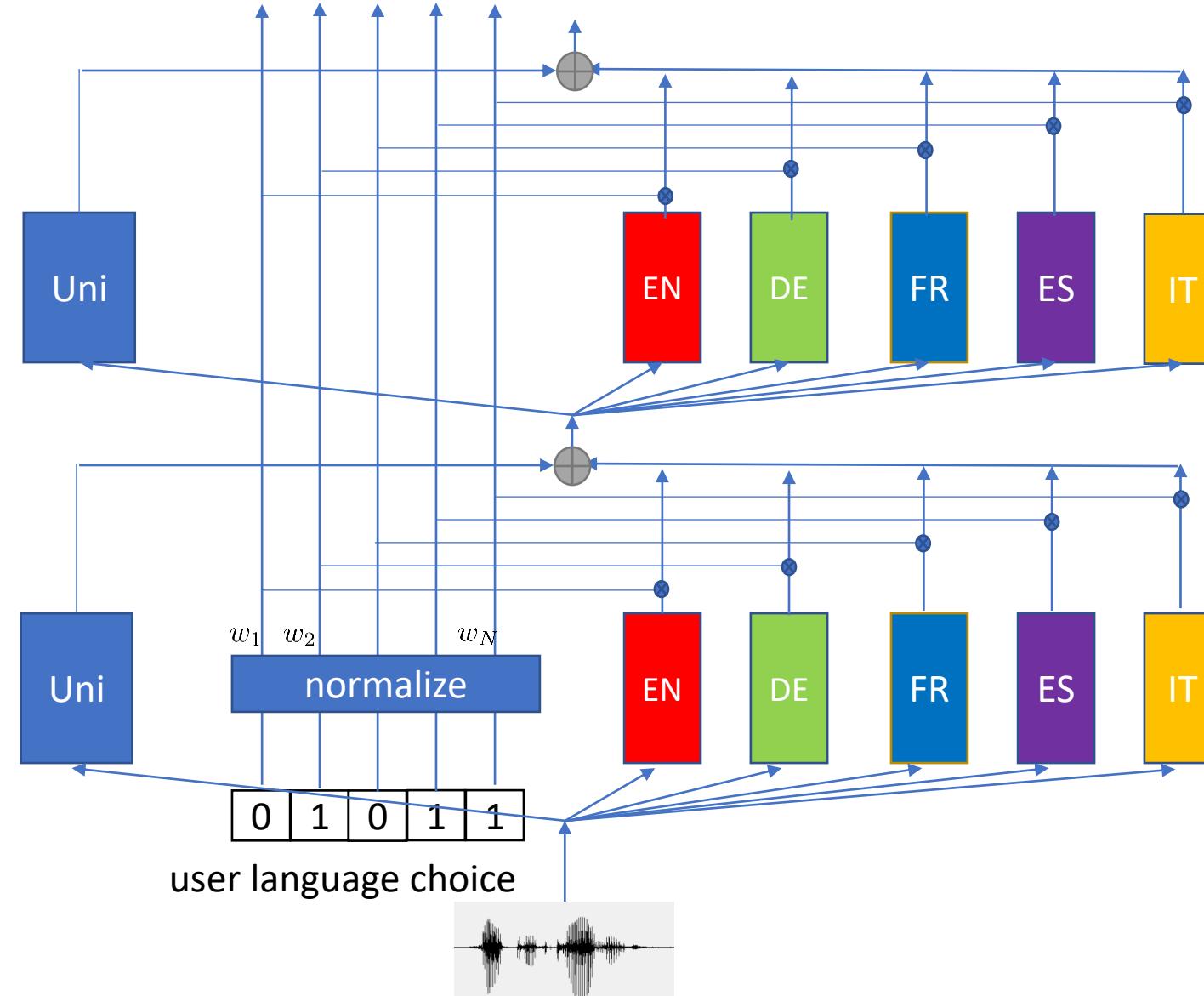
CMM

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



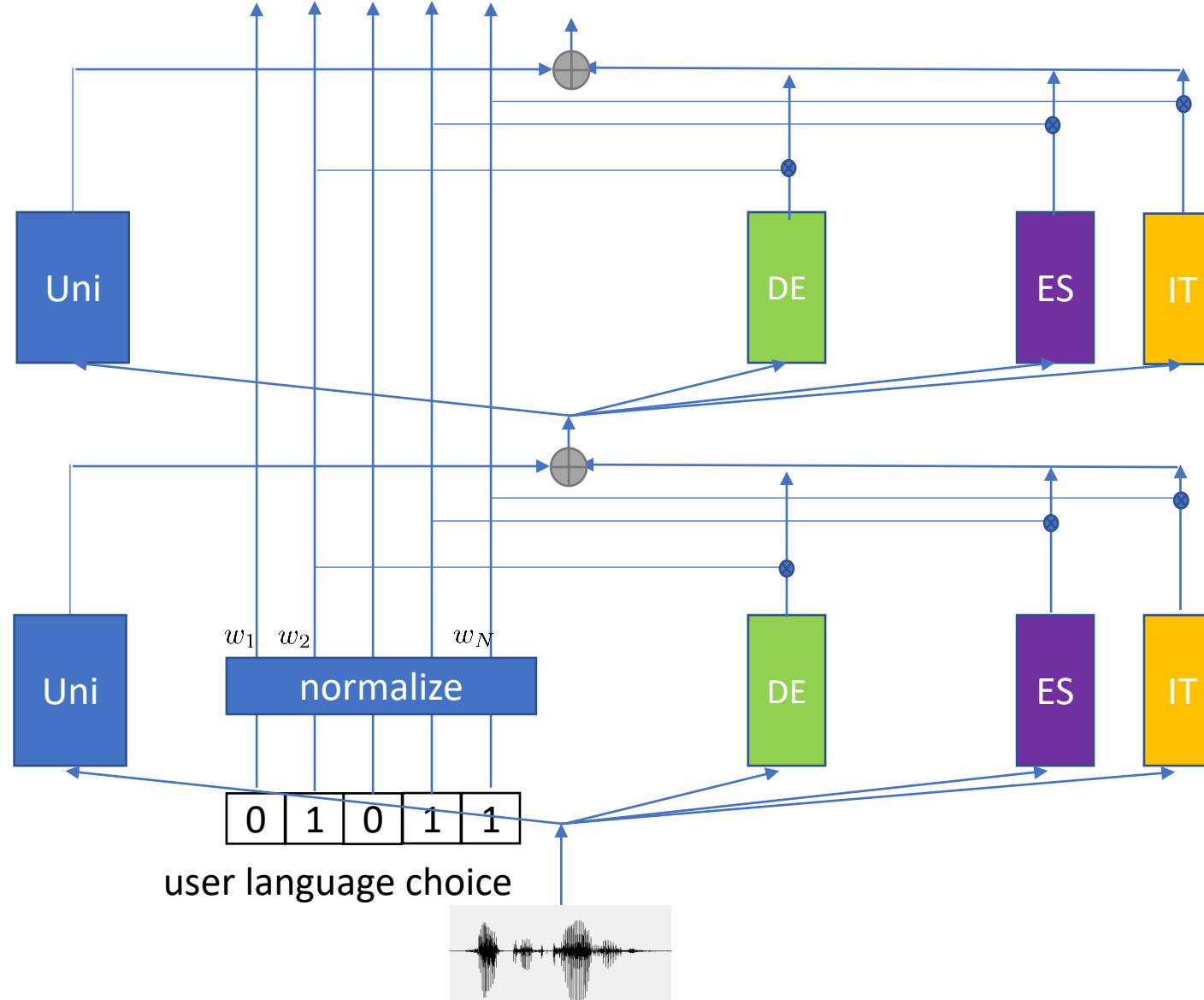
CMM

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



CMM

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





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.
- Two research directions:
 - Domain adaptation
 - Speech/Text joint training

Domain Adaptation

- The biggest challenge: not easy to get enough paired speech-text data in the new domain.
- Solution: utilize the new-domain text data.
 - LM fusion: fusing E2E models with an external LM trained with the new-domain text data.
 - Adaptation with text data: directly adjusting modules in E2E models
 - Adaptation with augmented audio: generating audio from the text to form the paired speech-text data to adjust E2E models.

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

- 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 et al., “On using monolingual corpora in neural machine translation,” arXiv preprint, 2015.

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

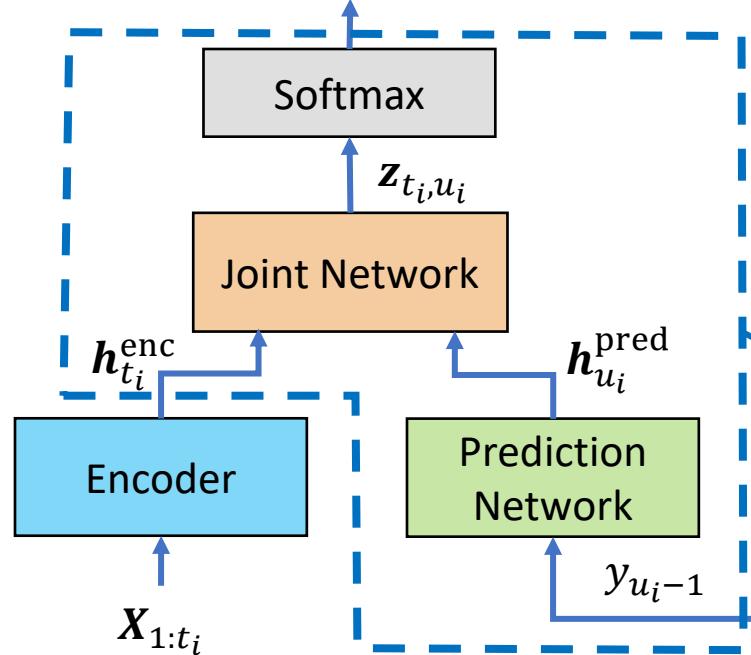
Variani et al. “Hybrid autoregressive transducer (HAT),” in Proc. ICASSP, 2020.

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

Internal LM Estimation

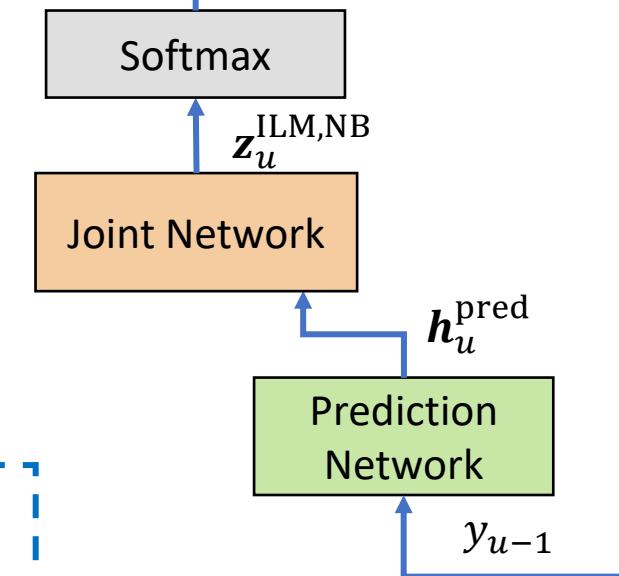
► RNN-T

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



► Internal LM estimation of RNN-T

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

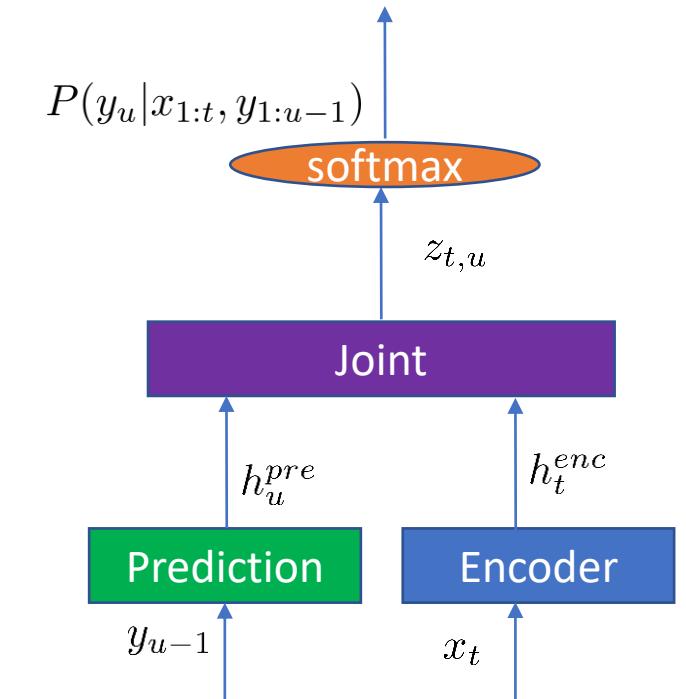


- Internal LM probability

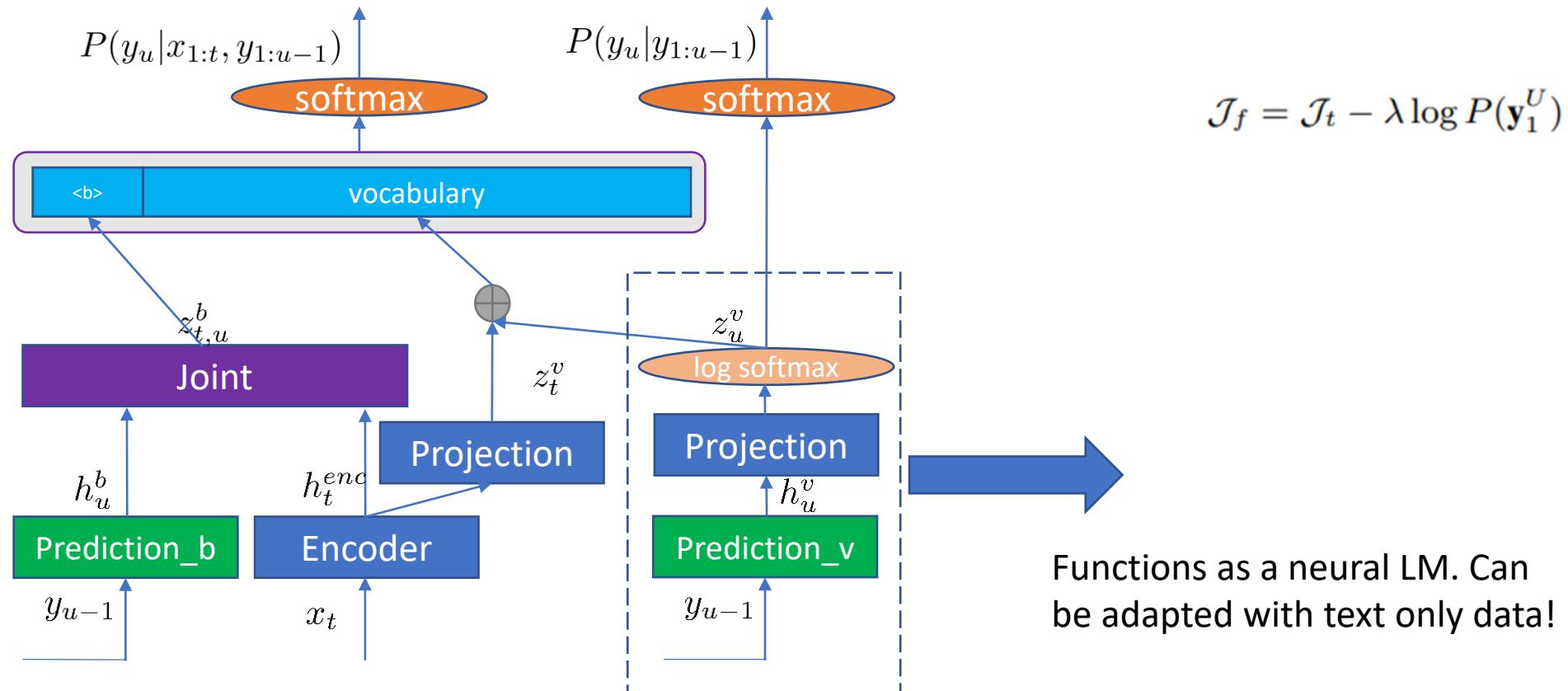
➤ The output of the **acoustically-conditioned LM** after removing the contribution of the encoder

Is the Prediction Network a LM?

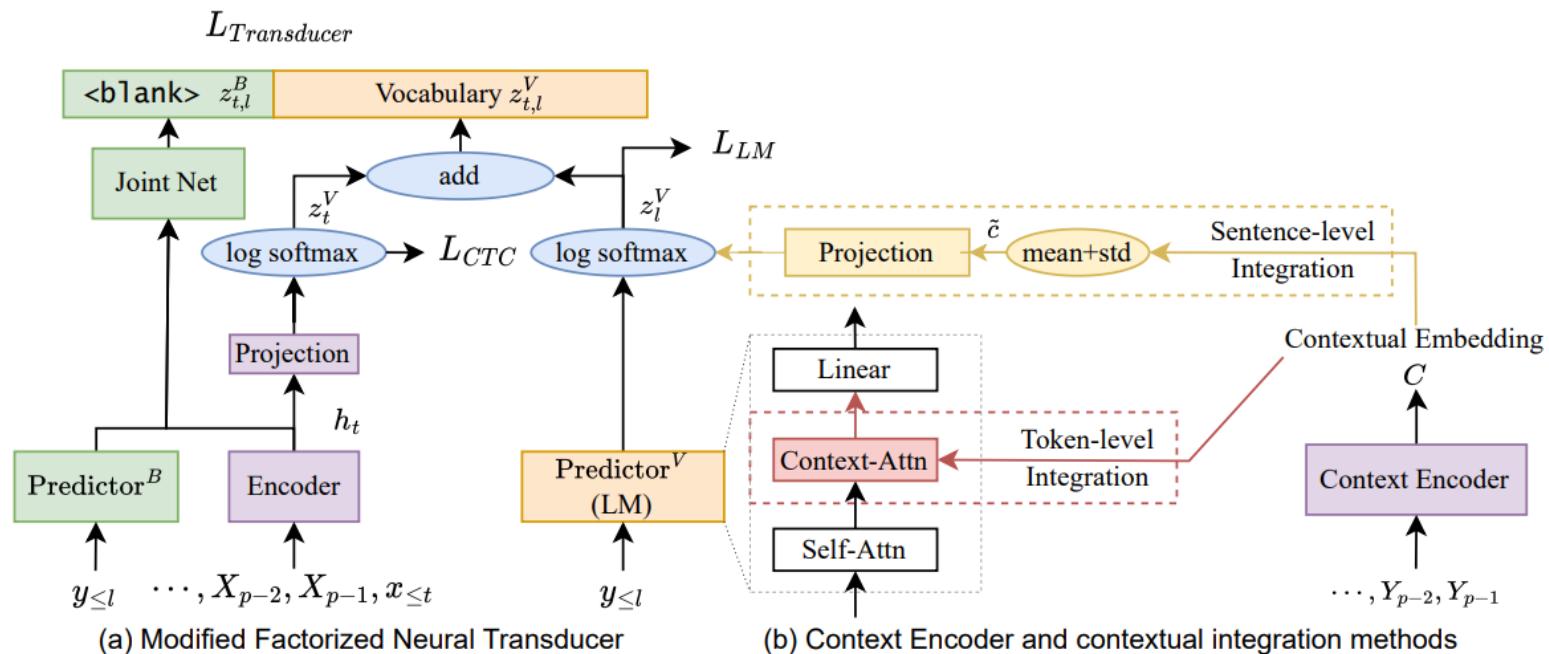
- If the prediction network in RNN-T is a LM, we can use new-domain text to adapt it.
- However, it does not fully function as a LM because it needs to predict both labels and blank.



Factorized Neural Transducer



LongFNT-Text: Explore Long-form Transcriptions



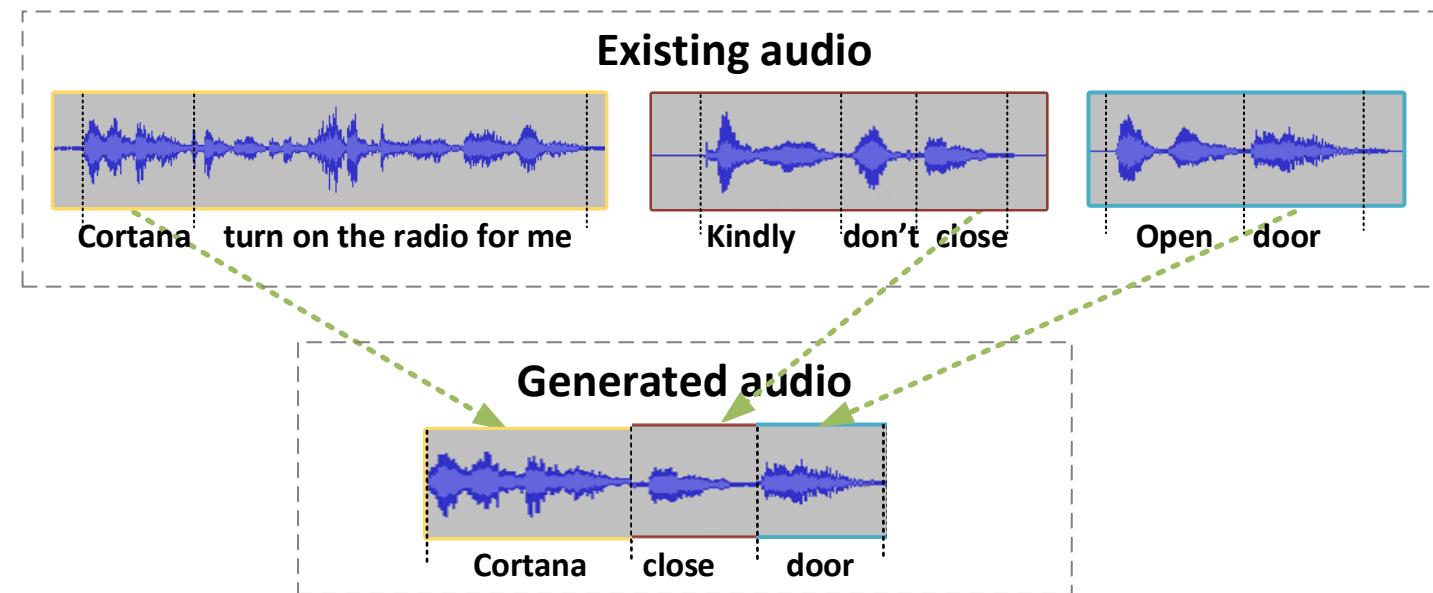
- Inspiration: FNT architecture explicitly separated out the LM part
- Two Level long-form contextual integrations:
 - sentence-level (statistical historical information)
 - token-level (using contextual attention)
- Long-form features are extracted using Context Encoder

TTS for Domain Adaptation

- Adapt E2E models with the synthesized speech generated from the new domain text.
- Drawbacks:
 - TTS speech is different from the real speech. It sometimes also degrades the recognition accuracy on real speech.
 - The speaker variation in the TTS data is far less than that in the large-scale ASR training data.
 - The cost of training a multi-speaker TTS model and the generation of synthesized speech from the model is large.

Data Splicing for Domain Adaptation

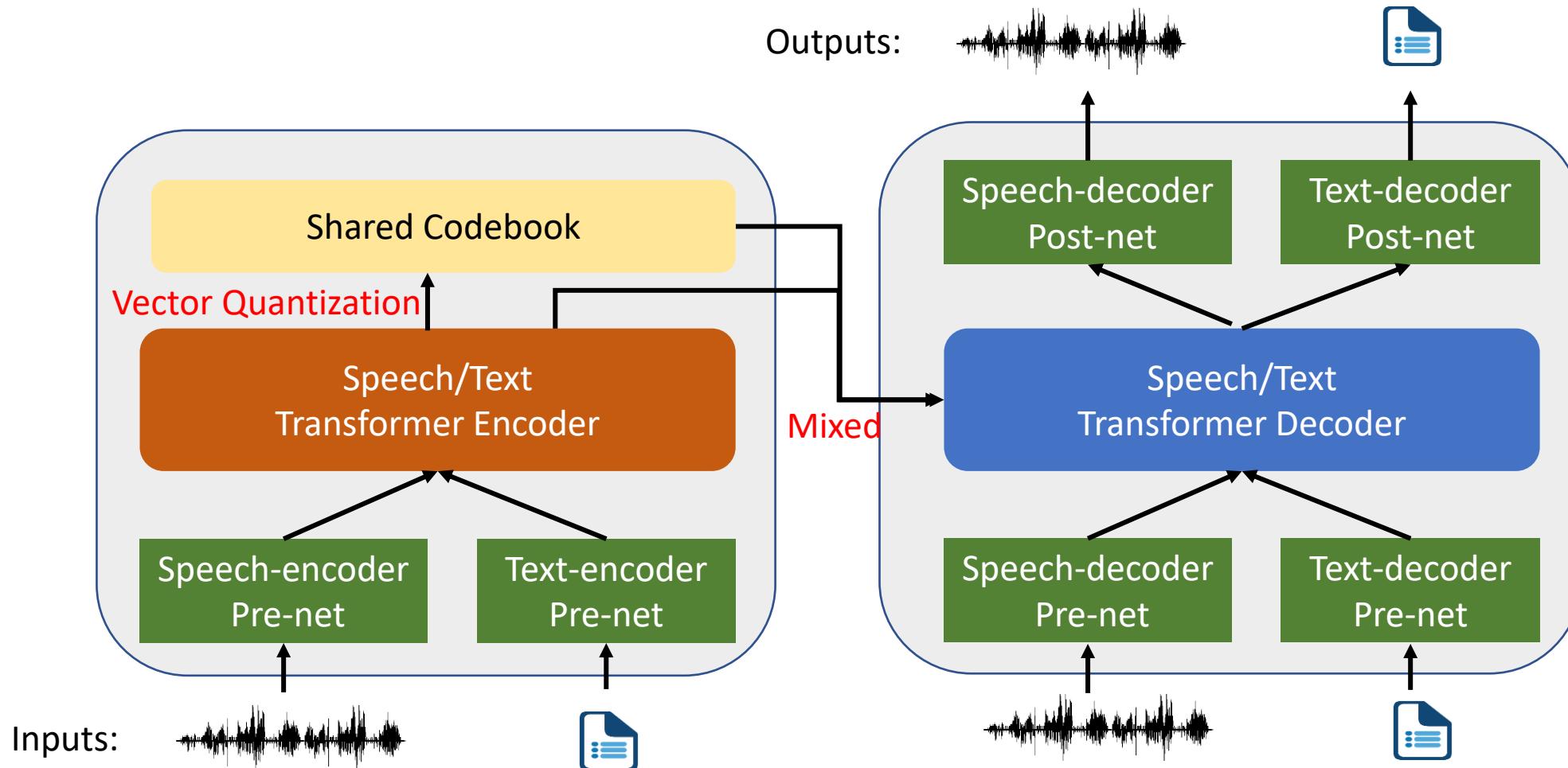
- Generate new audio from original ASR training data.



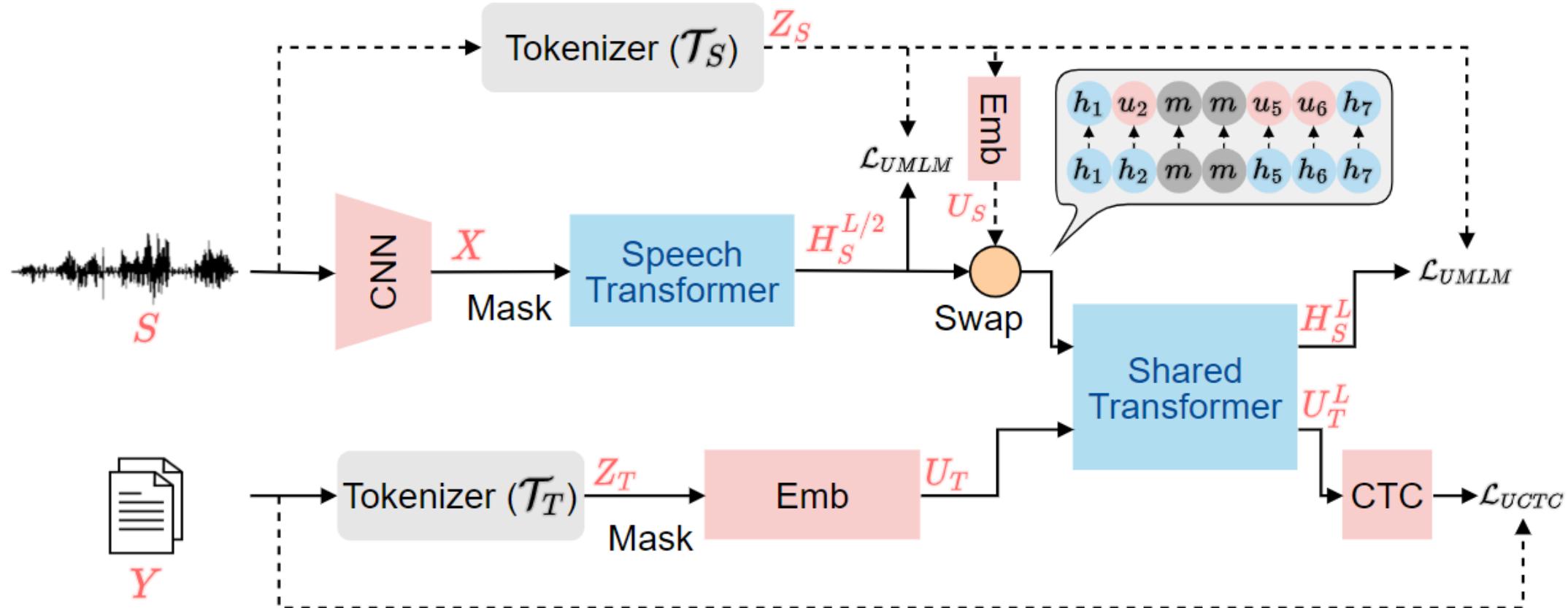
Speech/Text Joint Training

- Leverage large amount of unpaired text data
- Learn the unified representation for speech and text

SpeechT5

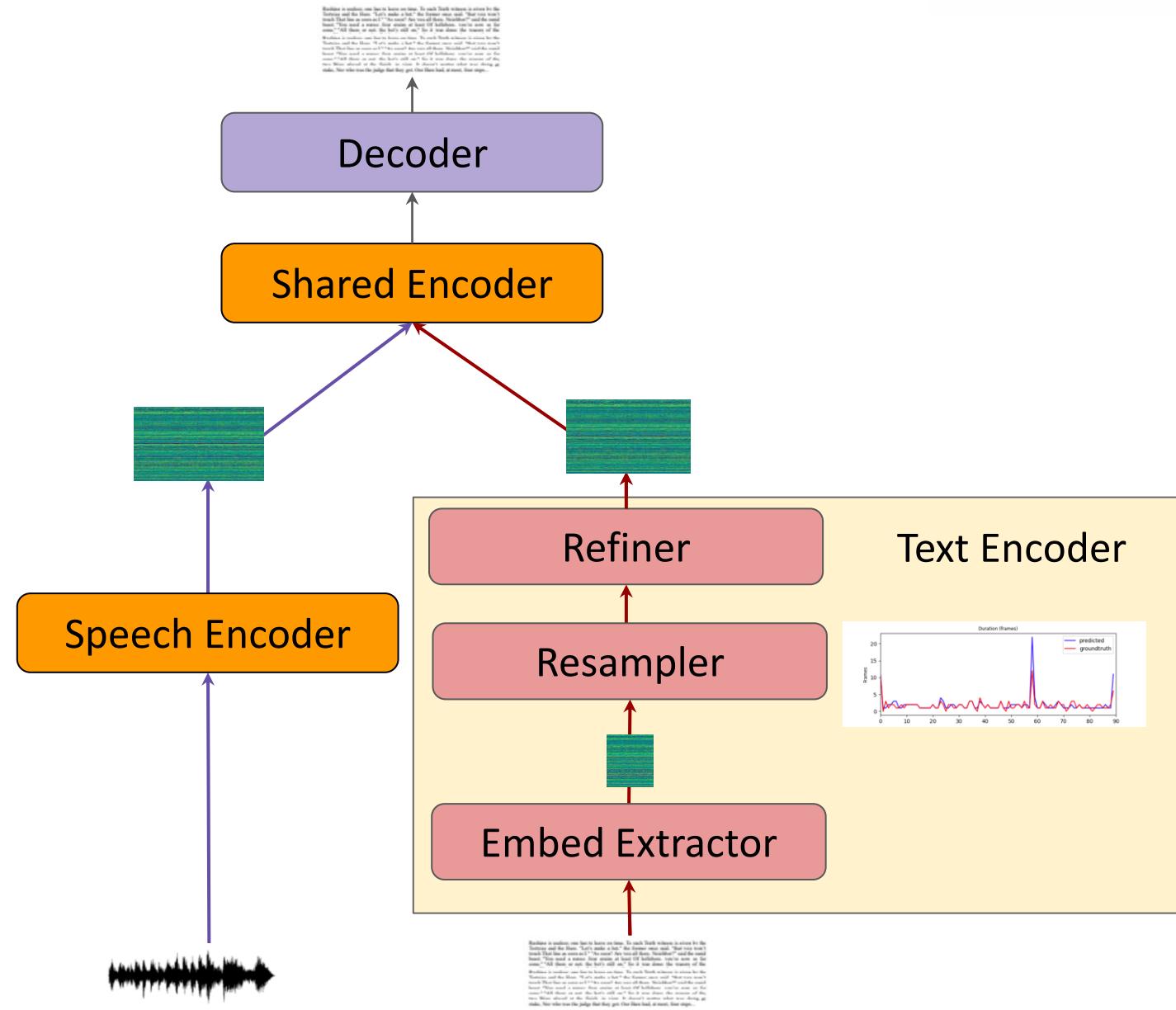


SpeechLM



Maestro

- Inject text representations and match the two modalities



Slides credit to Zhehuai Chen

Chen et al. "MAESTRO: Matched Speech Text Representations through Modality Matching." *Proc. Interspeech*, 2022.

Multi-talker ASR

Slides credit to Naoyuki Kanda

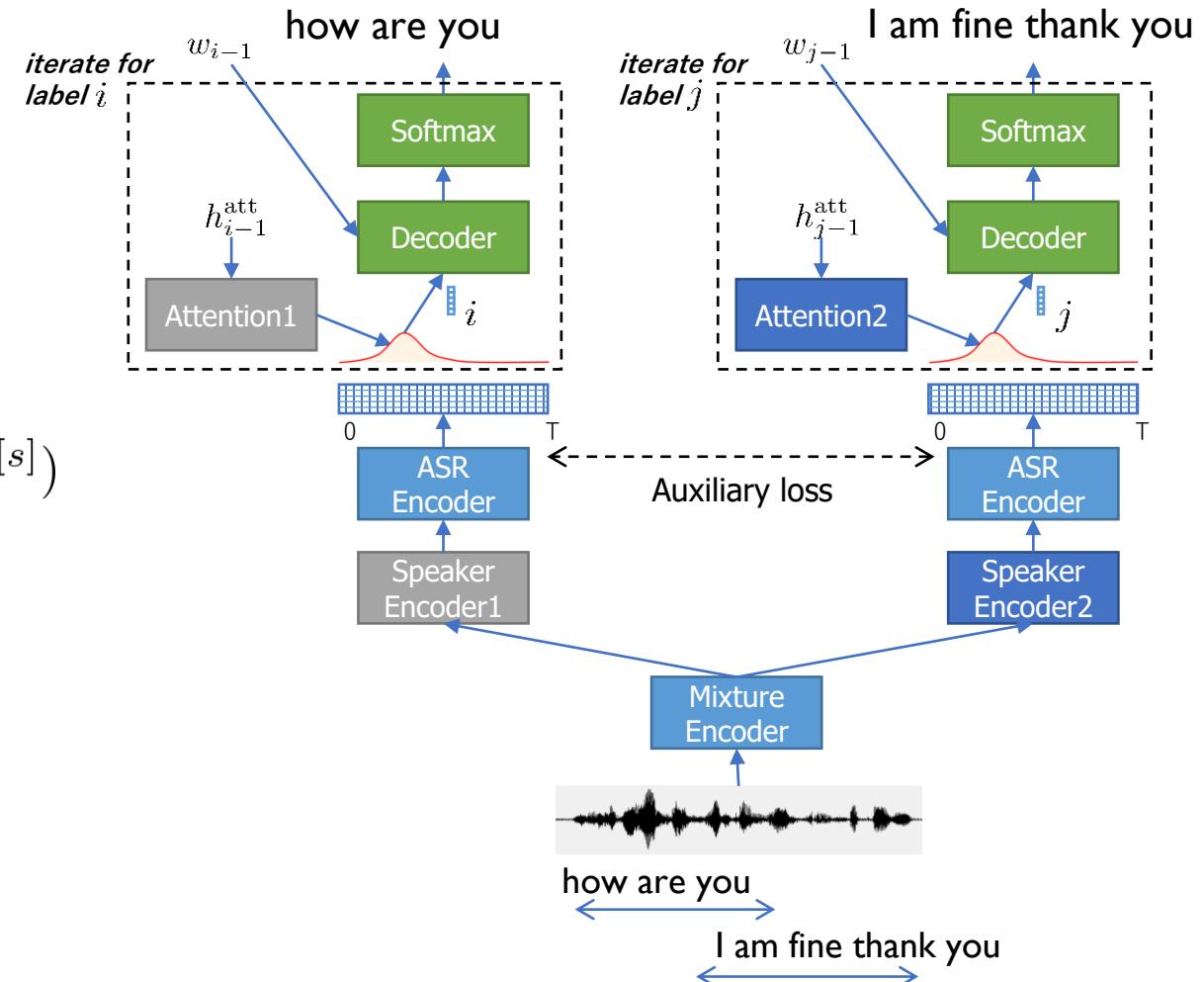
Multi-talker Models

- 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

Multi-talker AED model with PIT

- No need for noisy-clean audio pair for training.

$$L^{PIT} = \min_{\phi \in \Phi(1, \dots, S)} \sum_{s=1}^S CE(\mathbf{y}^s, \mathbf{r}^{\phi[s]})$$



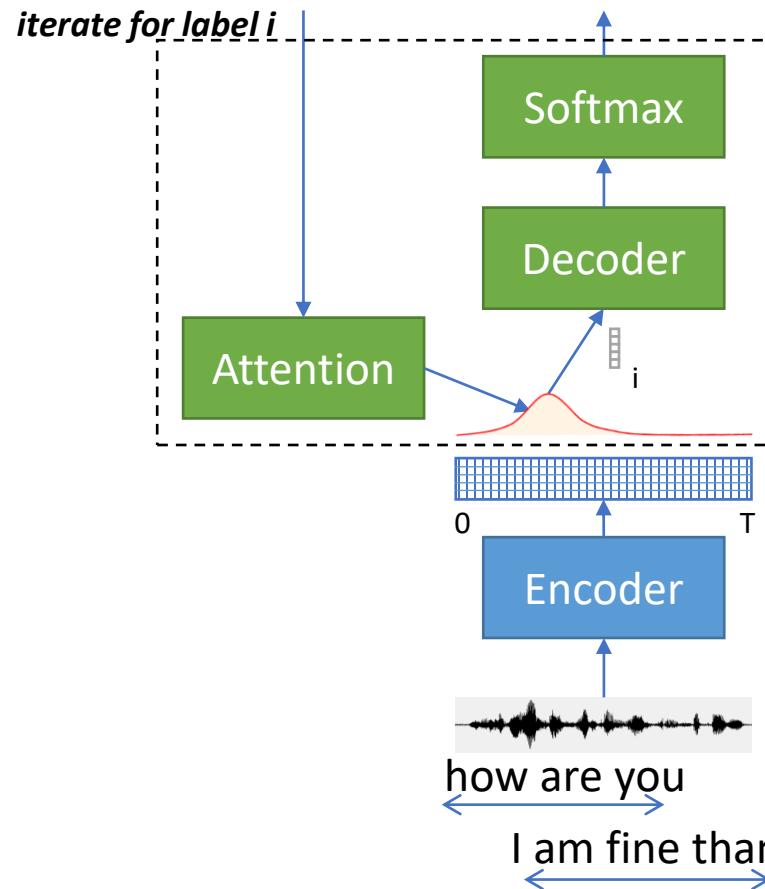
Seki et al., "A Purely End-to-end System for Multi-speaker Speech Recognition," ACL, 2018

Chang et al., End-to-End Monaural Multi-speaker ASR System without pretraining, Proc. ICASSP 2019.

Serialized Output Training (SOT)

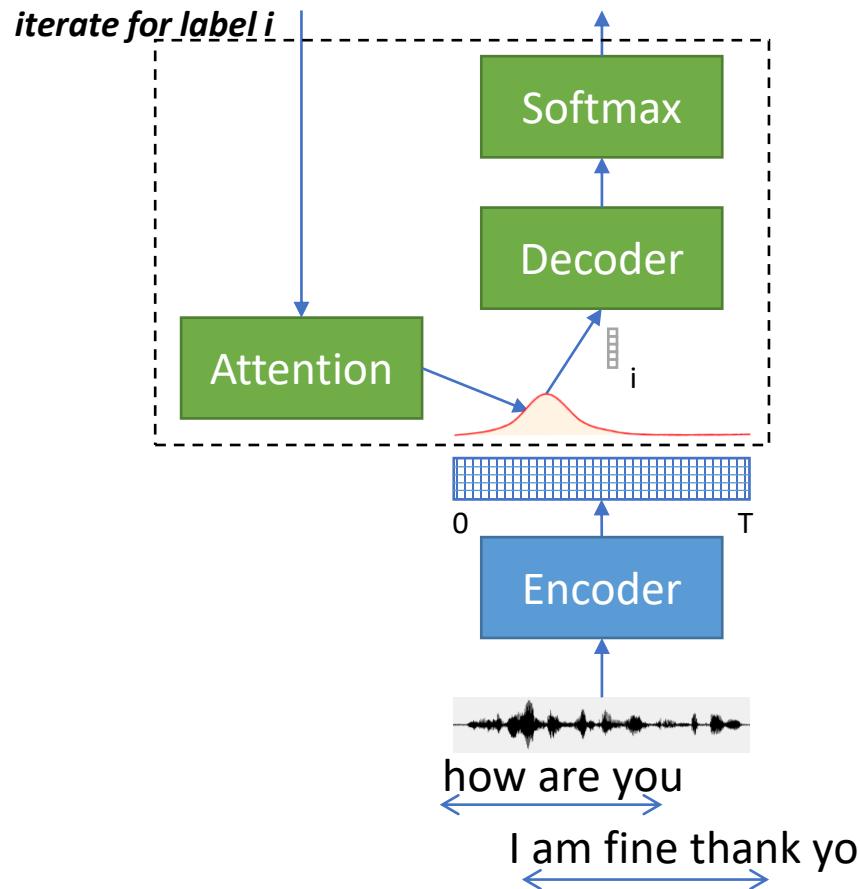


how are you <sc> I am fine thank you <eos>



Serialized Output Training (SOT)

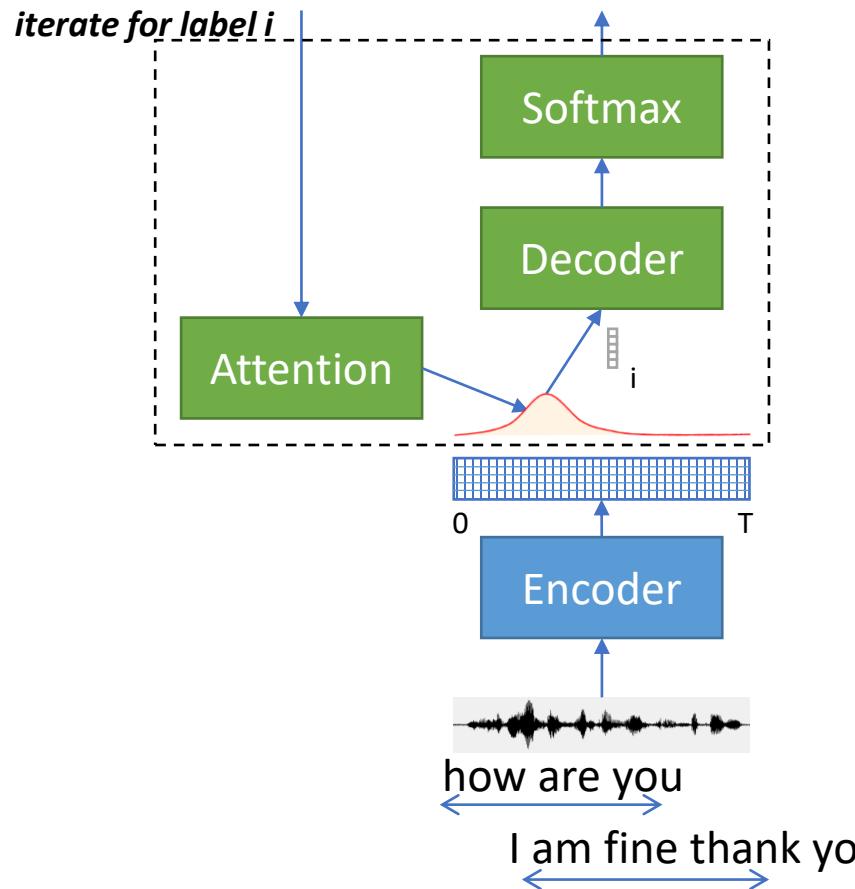
how are you <sc> I am fine thank you <eos>



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

Serialized Output Training (SOT)

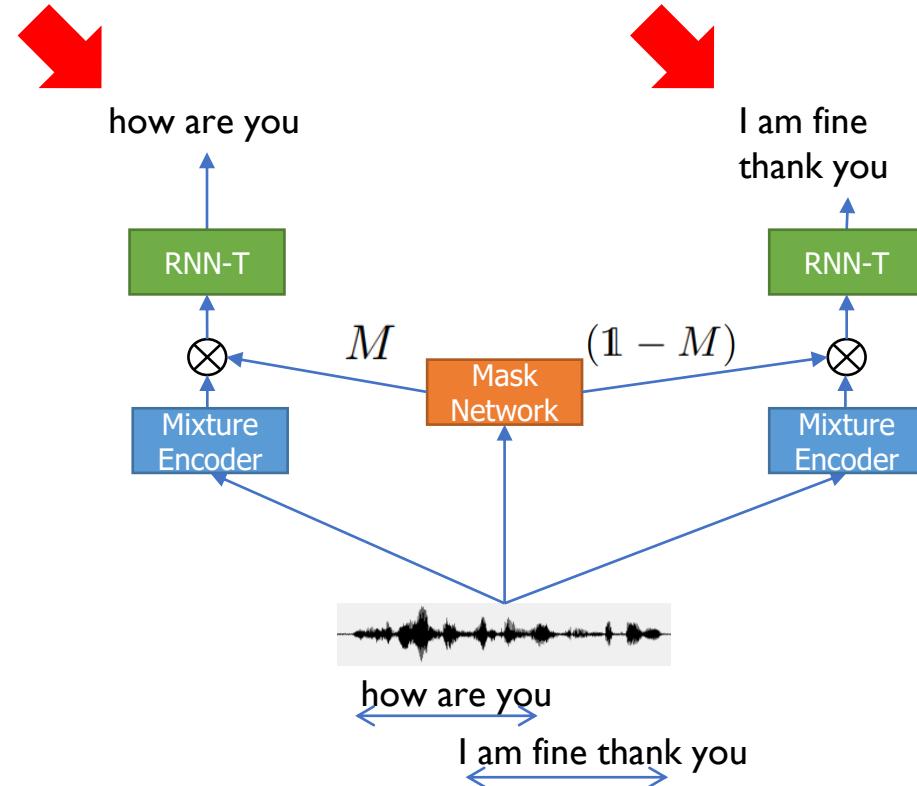
how are you <sc> I am fine thank you <eos>



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

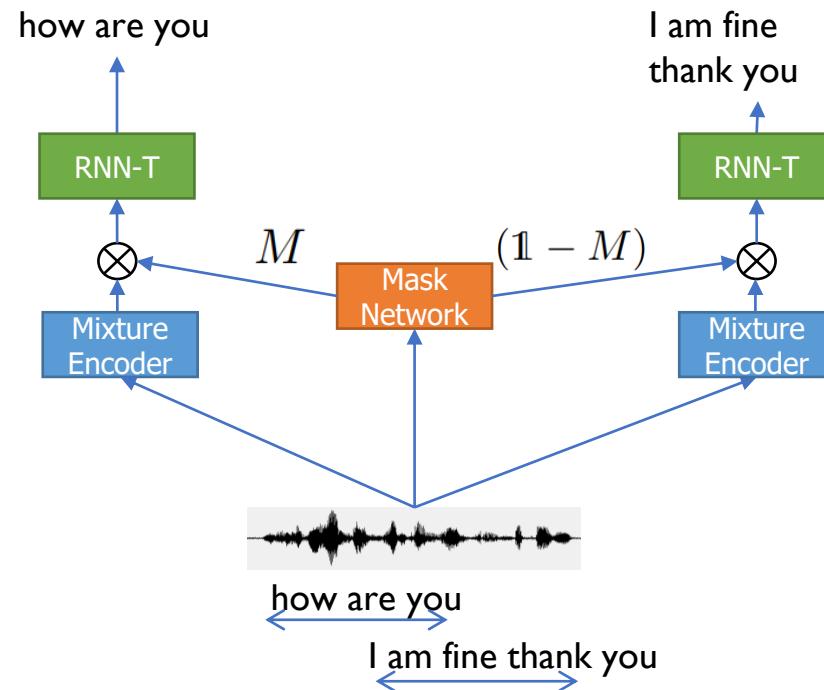
:(Only applicable for attention-based encoder decoder architecture
 → Only applicable for offline (i.e. non-streaming) inference

Streaming unmixing and recognition transducer (SURT)



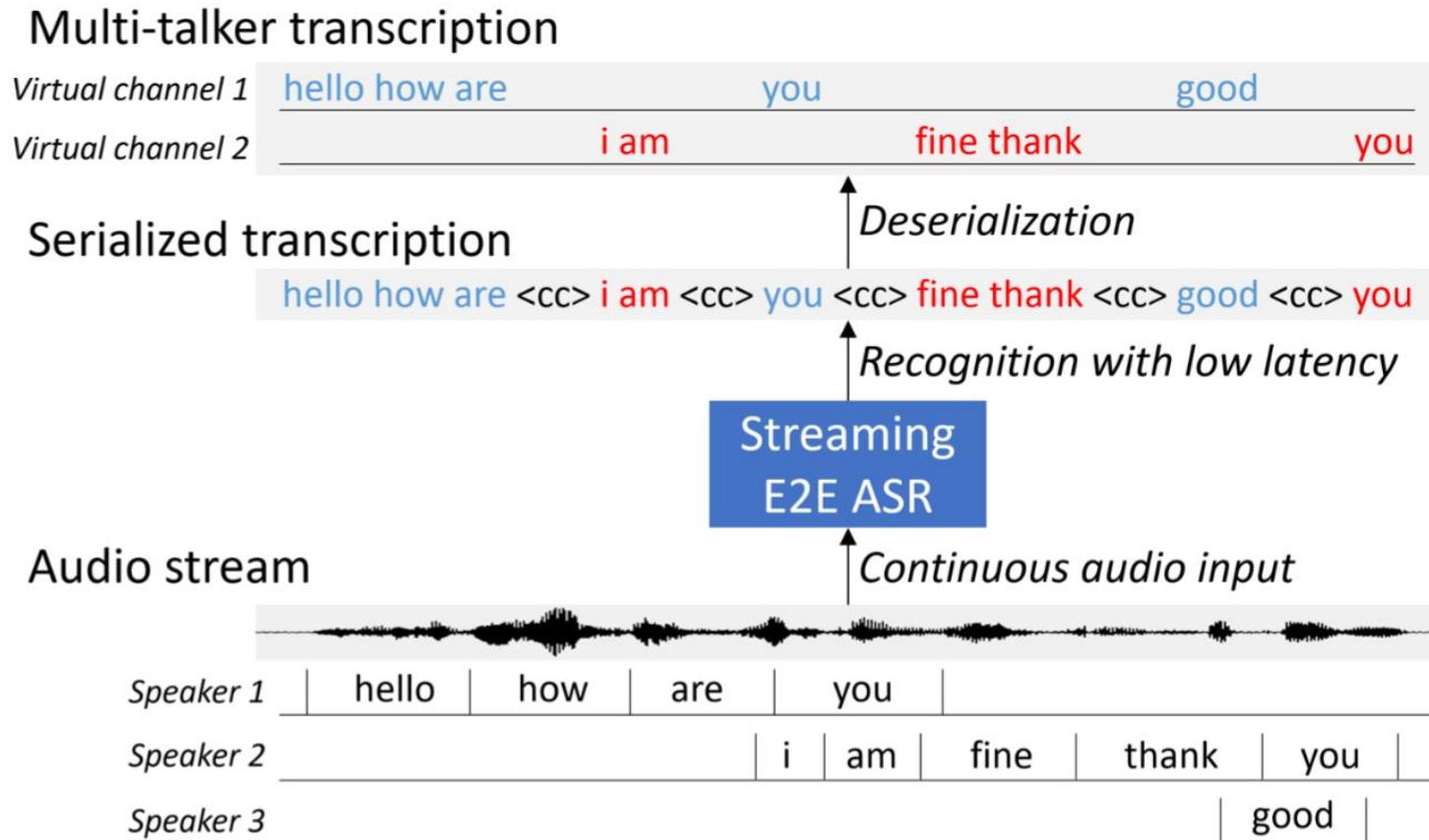
Heuristic Error Assignment Training:
order the label sequences based on the
utterance start time

Streaming unmixing and recognition transducer (SURT)



- 😊 • Can deal with overlapping speech
- 😊 • Streaming inference
- 😢 • Complicated
- 😢 • Accuracy was not very good
 - Duplicated hypotheses issue
 - No hypothesis issue

Token-level Serialized Output Training (t-SOT)



Token-level Serialized Output Training (t-SOT)

Multi-talker transcription

<i>Virtual channel 1</i>	hello how are	you	good	
<i>Virtual channel 2</i>		i am	fine thank	you

Serialized transcription

hello how are <cc> i am <cc> you <cc> fine thank <cc> good <cc> you

Deserialization

Recognition with low latency

Streaming
E2E ASR

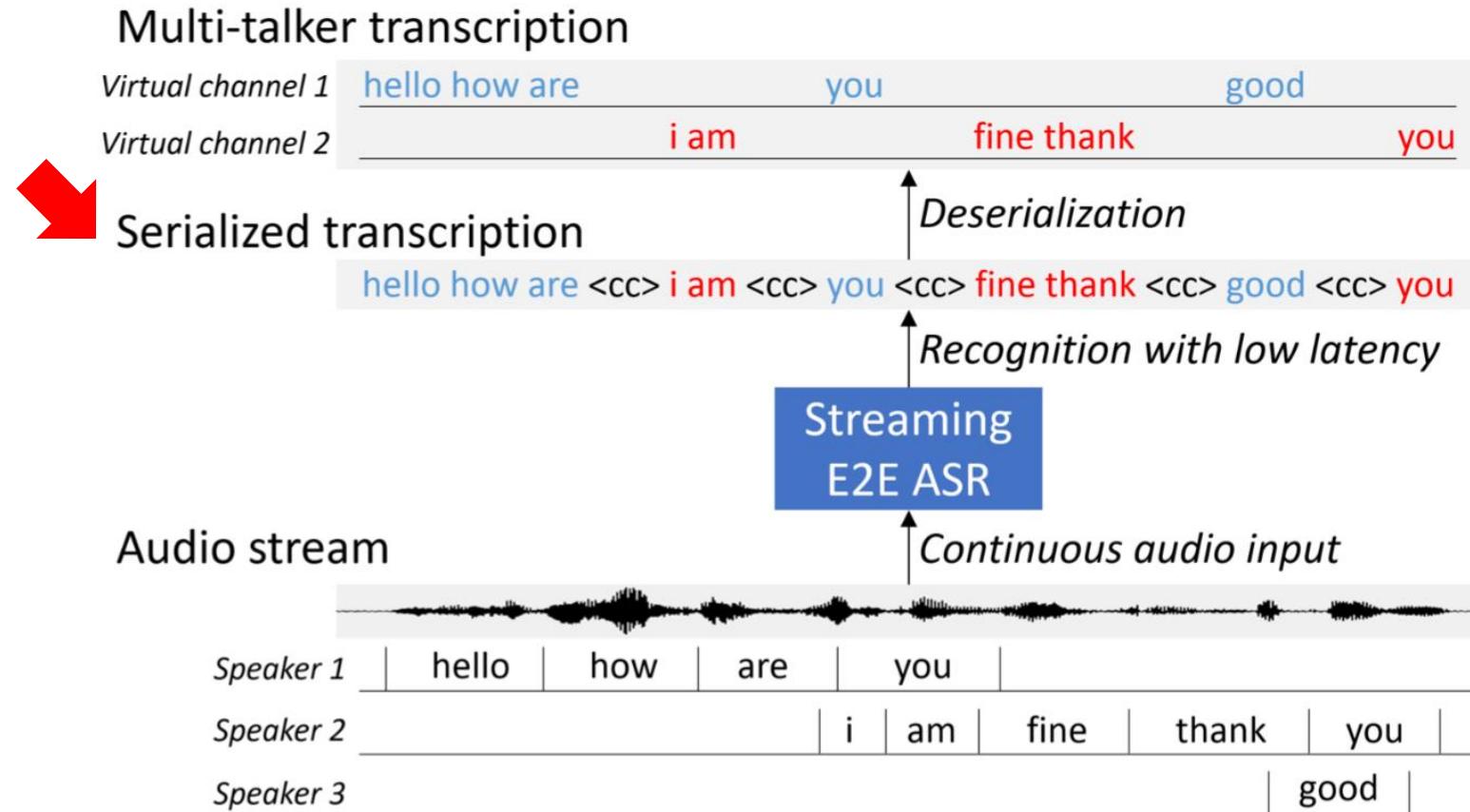
Audio stream



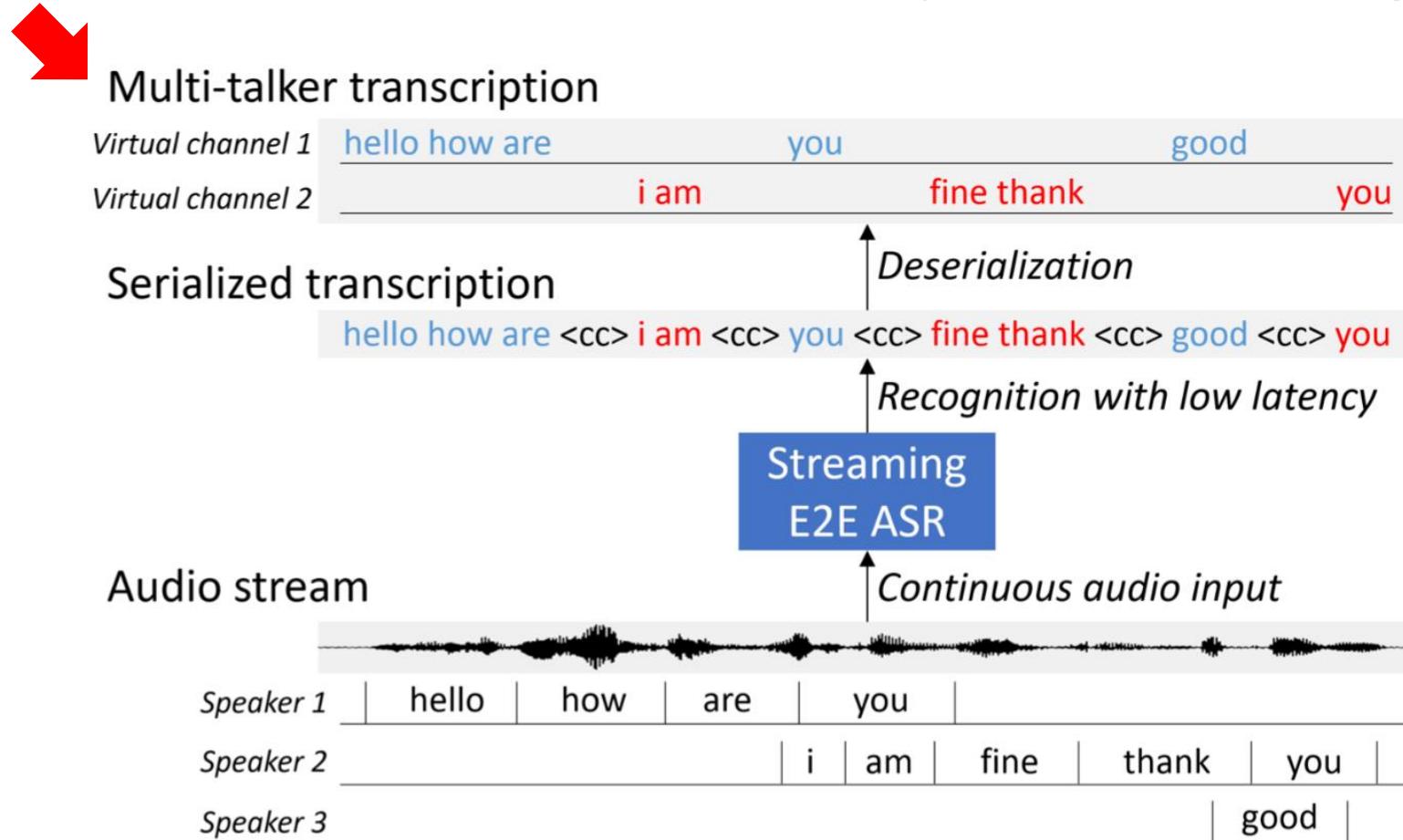
Continuous audio input

<i>Speaker 1</i>	hello	how	are	you	
<i>Speaker 2</i>			i	am	fine thank you
<i>Speaker 3</i>					good

Token-level Serialized Output Training (t-SOT)

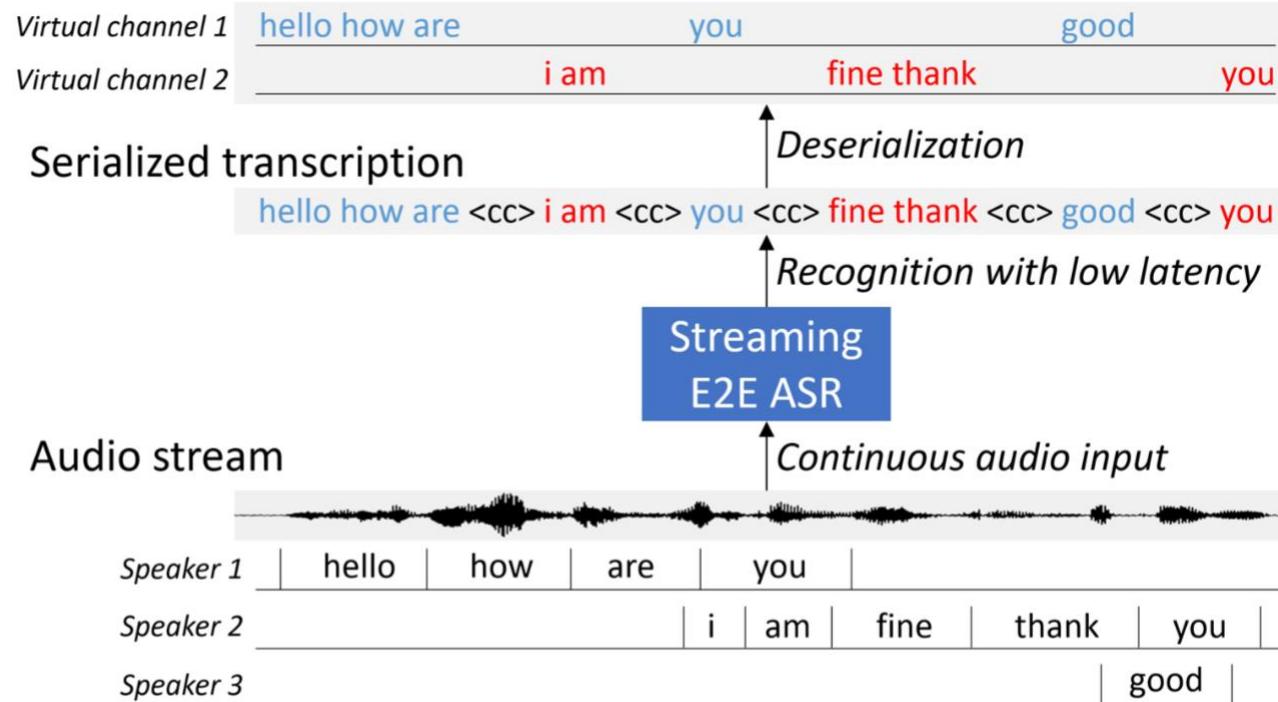


Token-level Serialized Output Training (t-SOT)



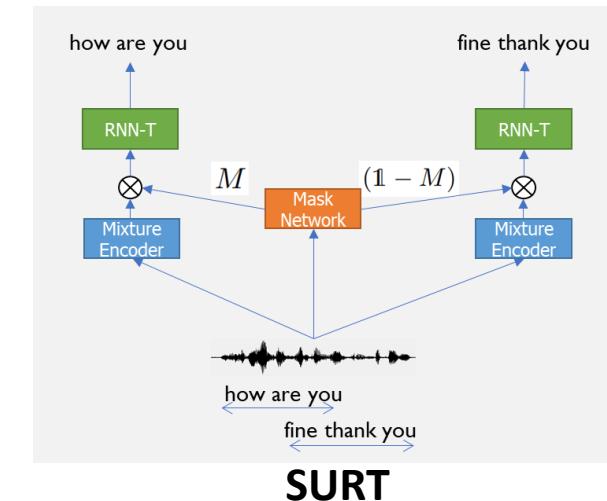
Token-level Serialized Output Training (t-SOT)

Multi-talker transcription



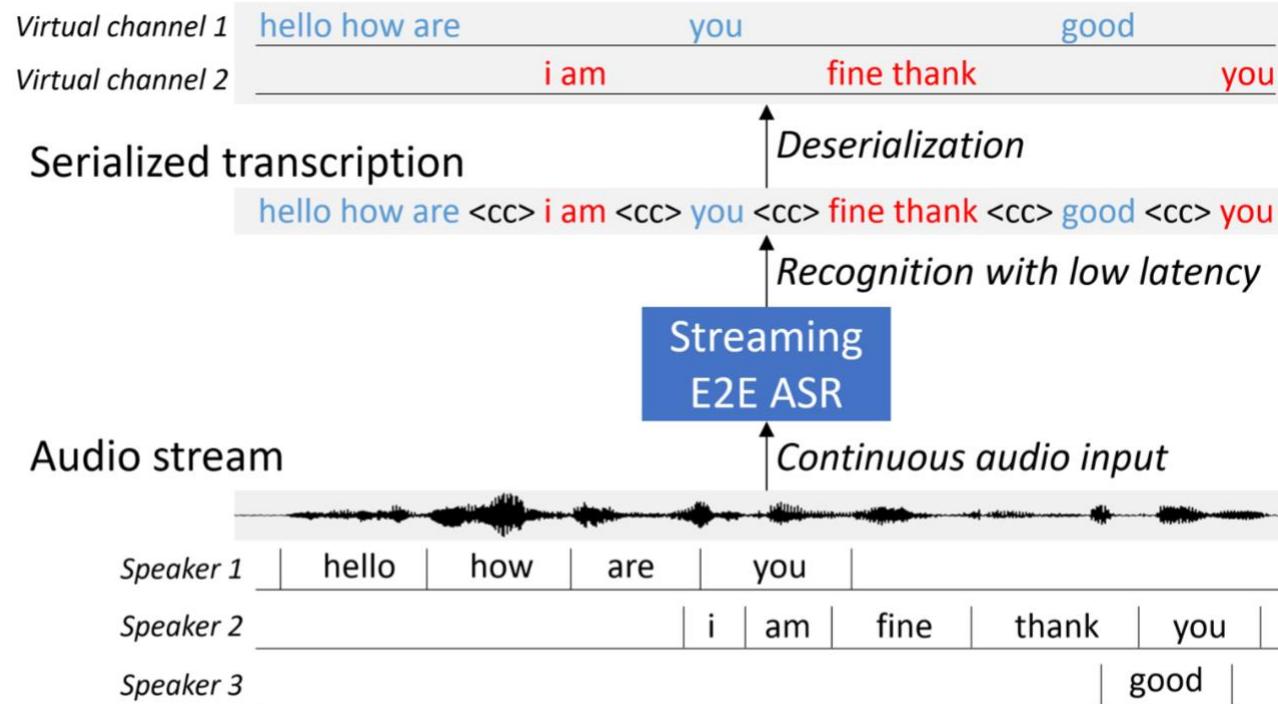
t-SOT vs. SURT

- t-SOT is simpler
- t-SOT requires less computation
- t-SOT is significantly better in accuracy



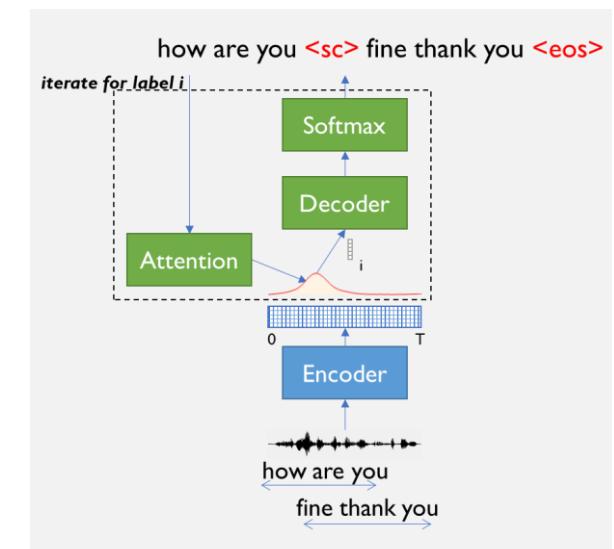
Token-level Serialized Output Training (t-SOT)

Multi-talker transcription

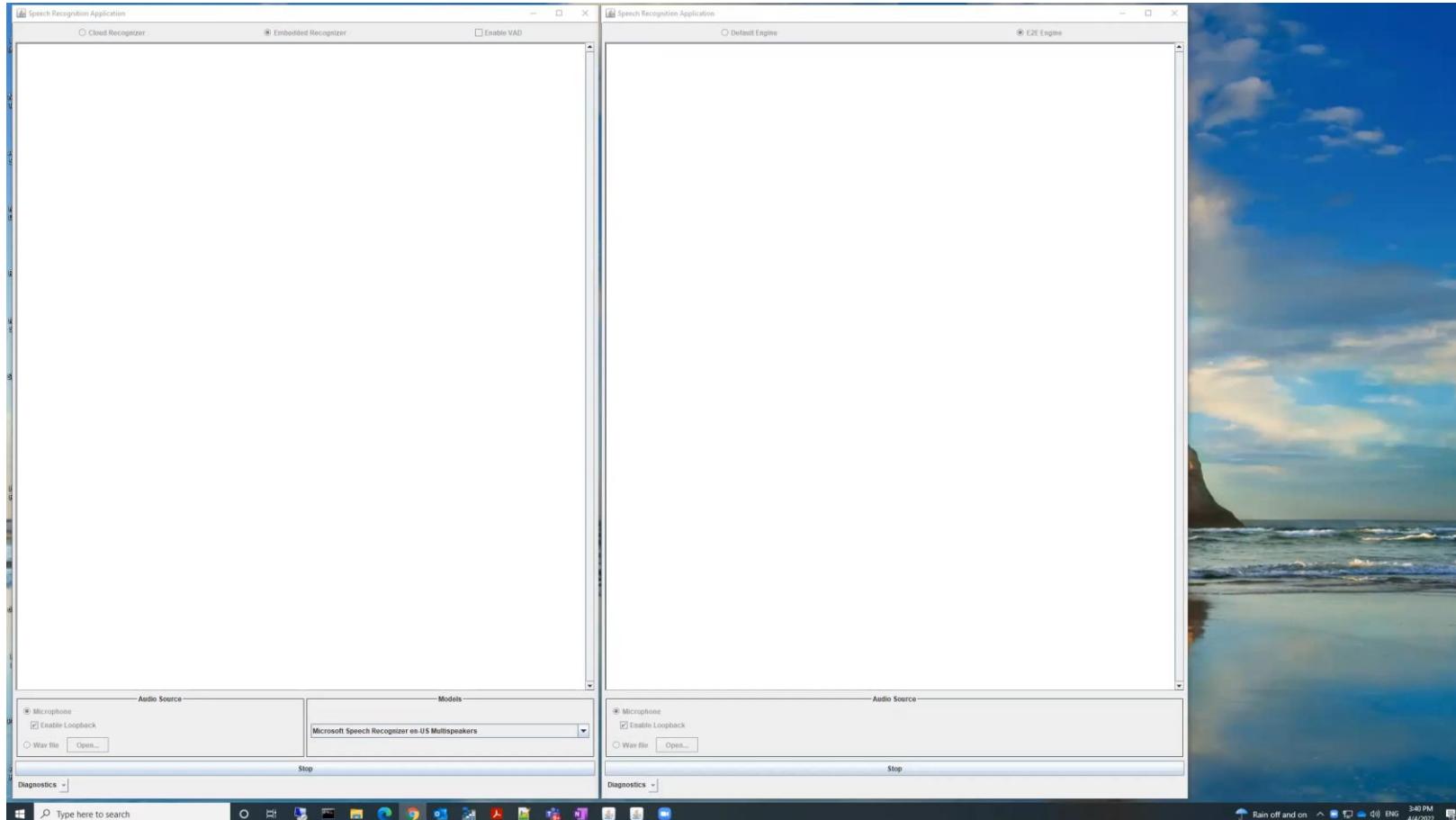


t-SOT vs. SOT

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



Multi-talker ASR Demo



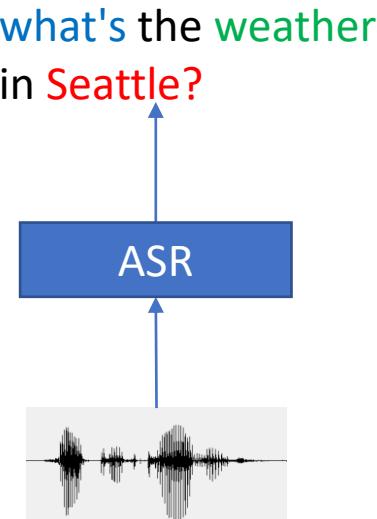
Simultaneous Speech Translation

Popular Simultaneous Speech Translation (ST) Methods

- Re-Translation: re-translate partial ASR results from beginning
 - Cost is very high because machine translation (MT) needs to be called multiple times
 - Stability is an issue because the outputs of different MT calls are independent
- Wait-K: start to translate ASR results after waiting for K words.
 - The read-write operation is interleaving, not flexible
 - K is pre-determined
- AED models for E2E ST
 - Streaming AED is still a challenge

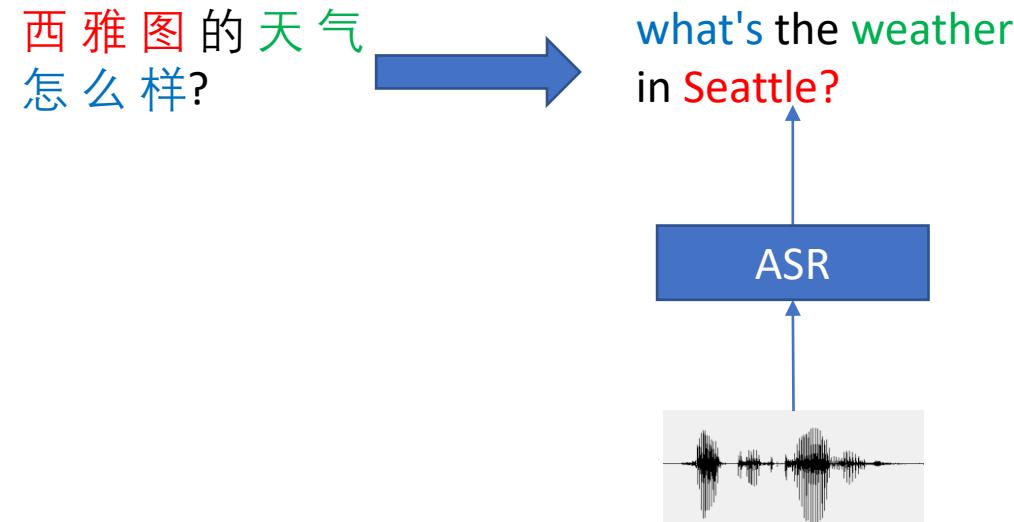
Can We Build a Simultaneous E2E ST System?

- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.



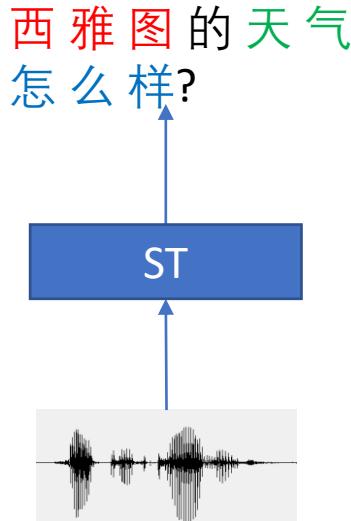
Can We Build a Simultaneous Direct ST System?

- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.

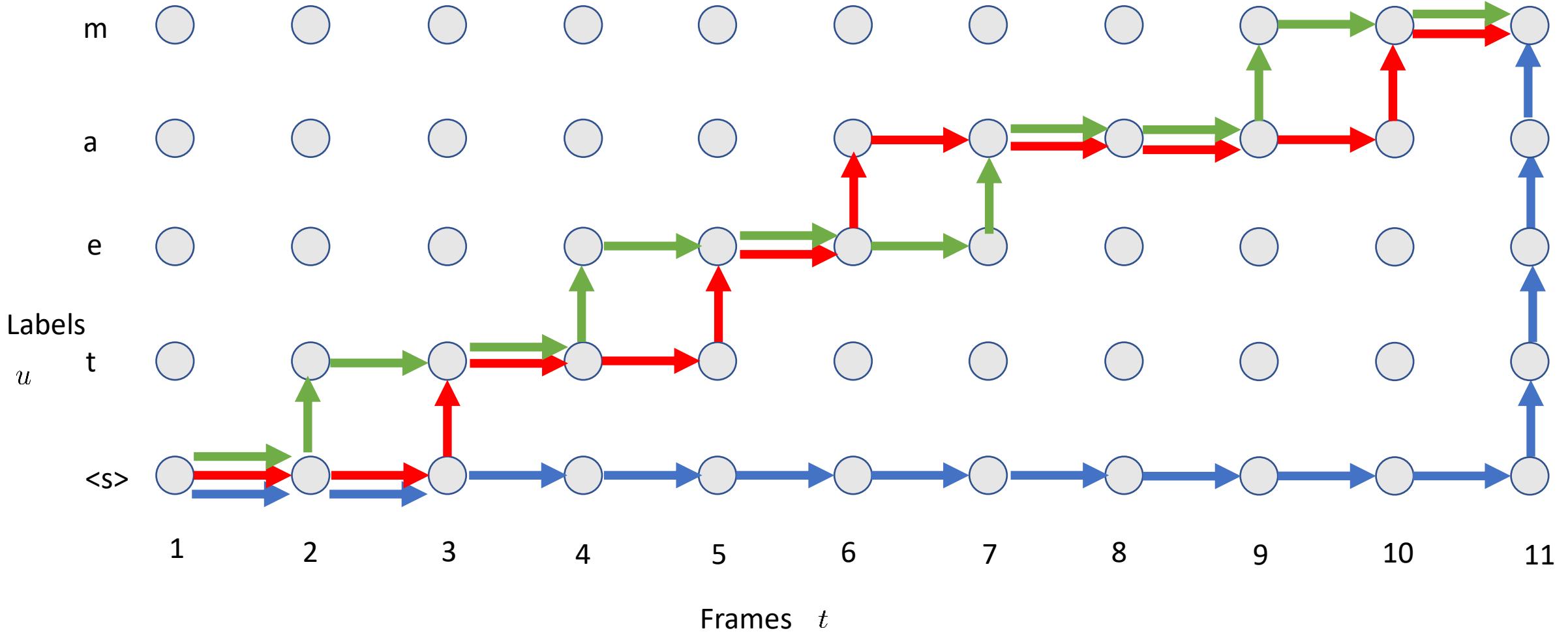


Can We Build a Simultaneous Direct ST System?

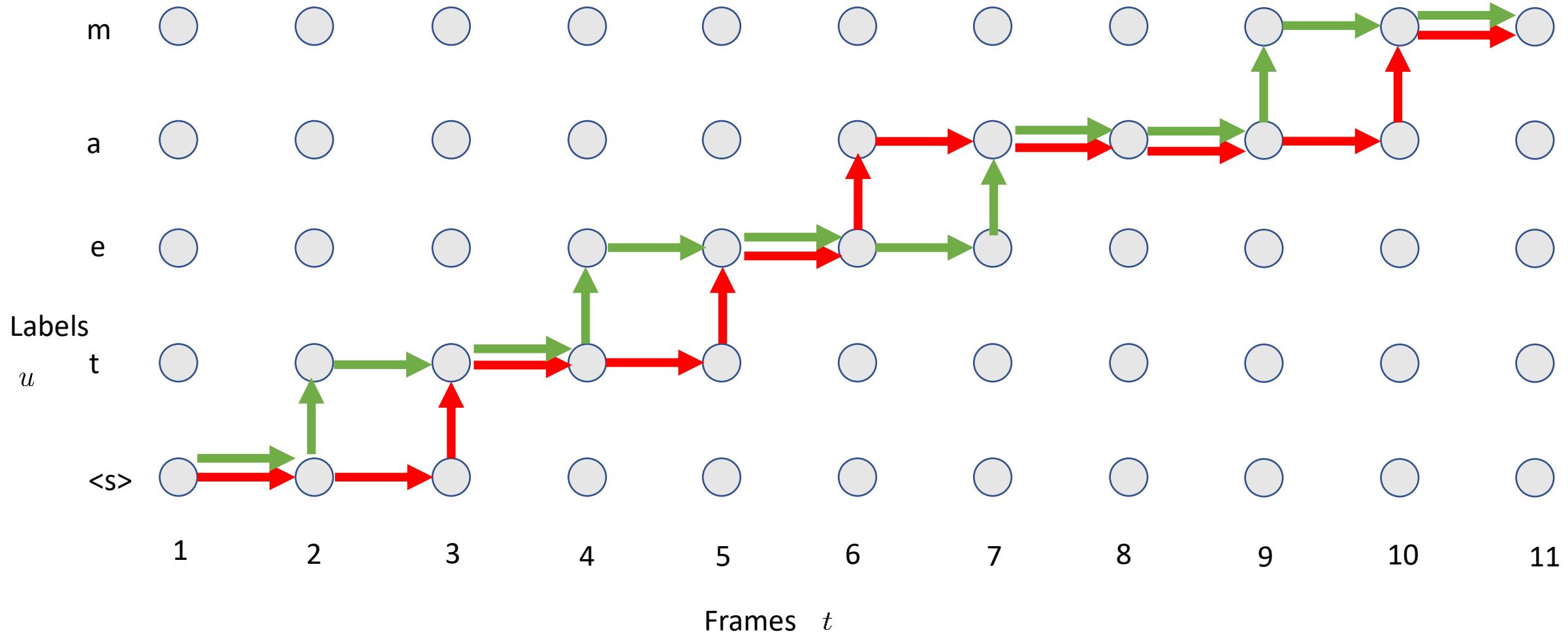
- Treating ST as an ASR problem – we already have the success in streaming E2E ASR.
- We directly use streaming Transformer Transducer to build streaming ST.



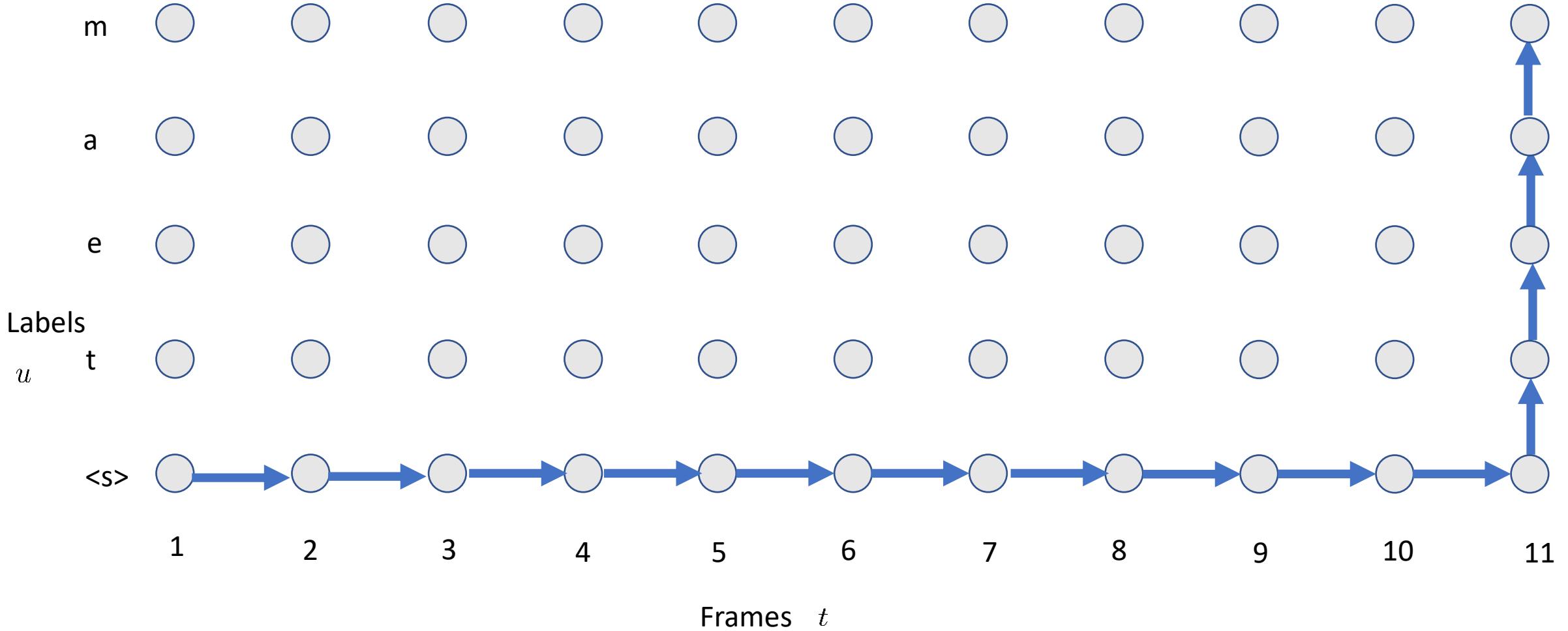
Flexible RNN-T Path



No Significant Word-Reordering



Word-Reordering at the End of Utterance

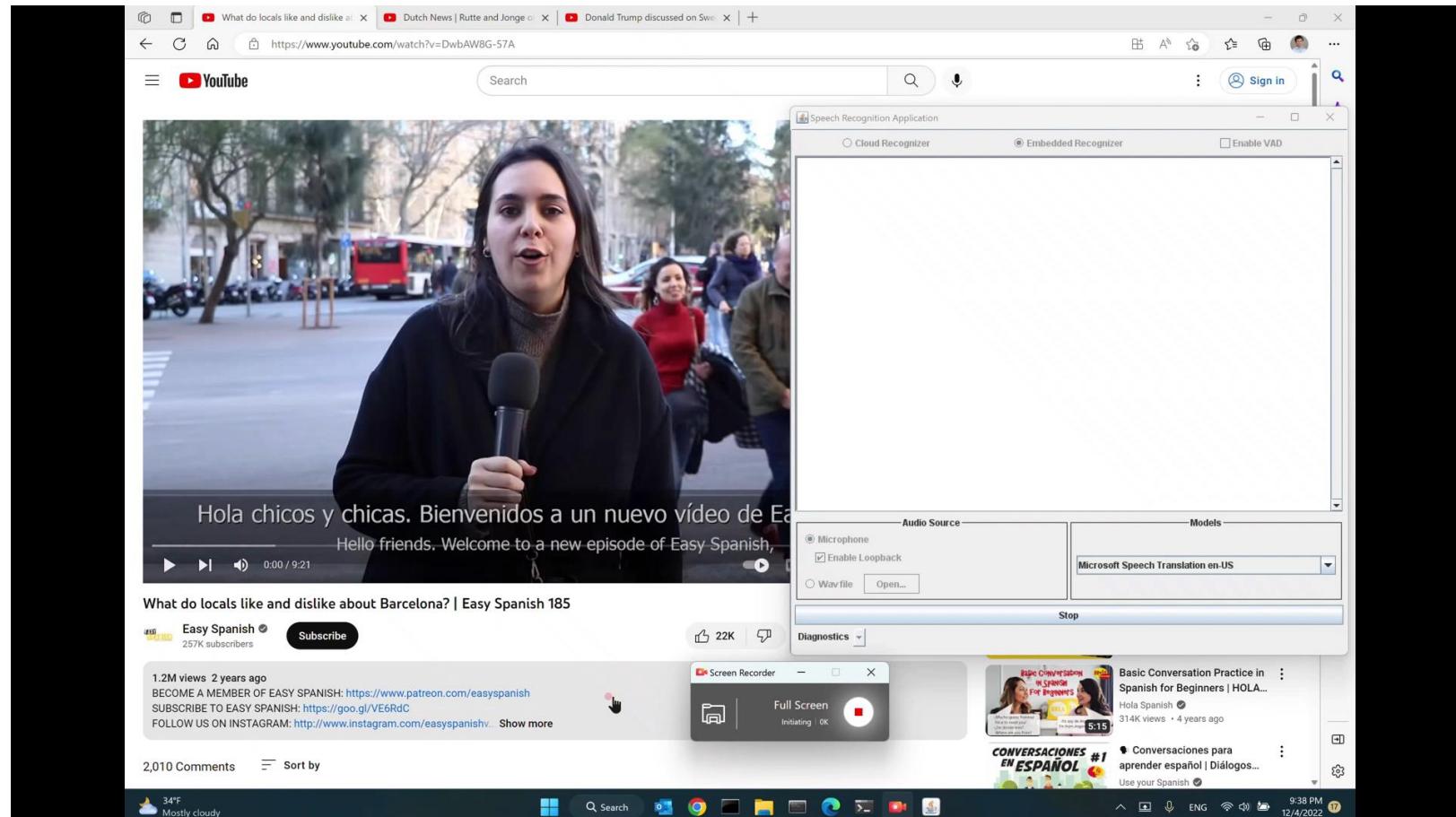


Streaming Multilingual Speech Model (SM²)

- 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.

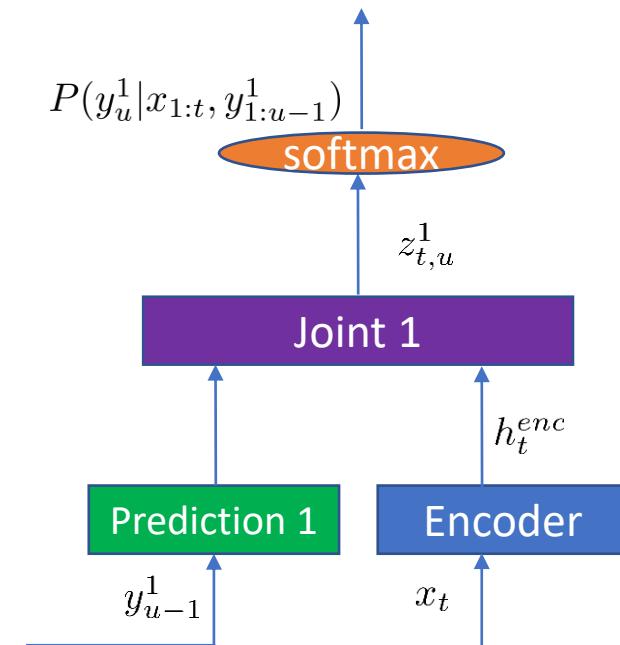
Xue et al. "A Weakly-Supervised Streaming Multilingual Speech Model with Truly Zero-Shot Capability." *arXiv preprint*, 2022.

SM² Trained with 25 Languages->English



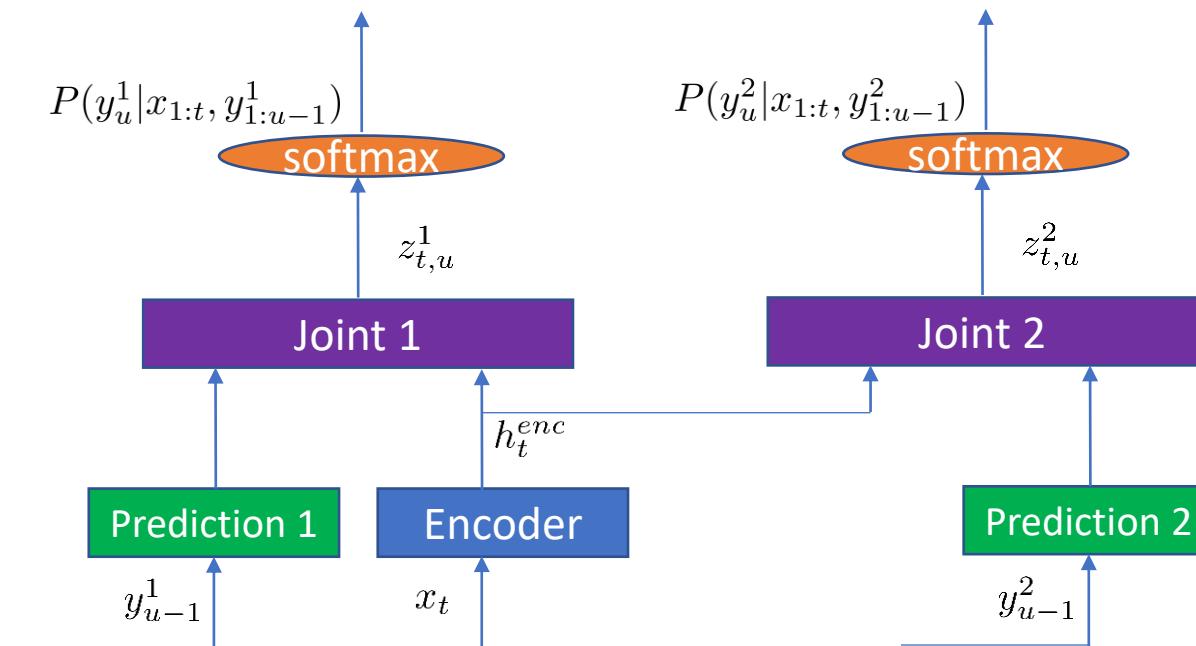
Language Expansion

- Every language has its own prediction and joint network, sharing the same encoder



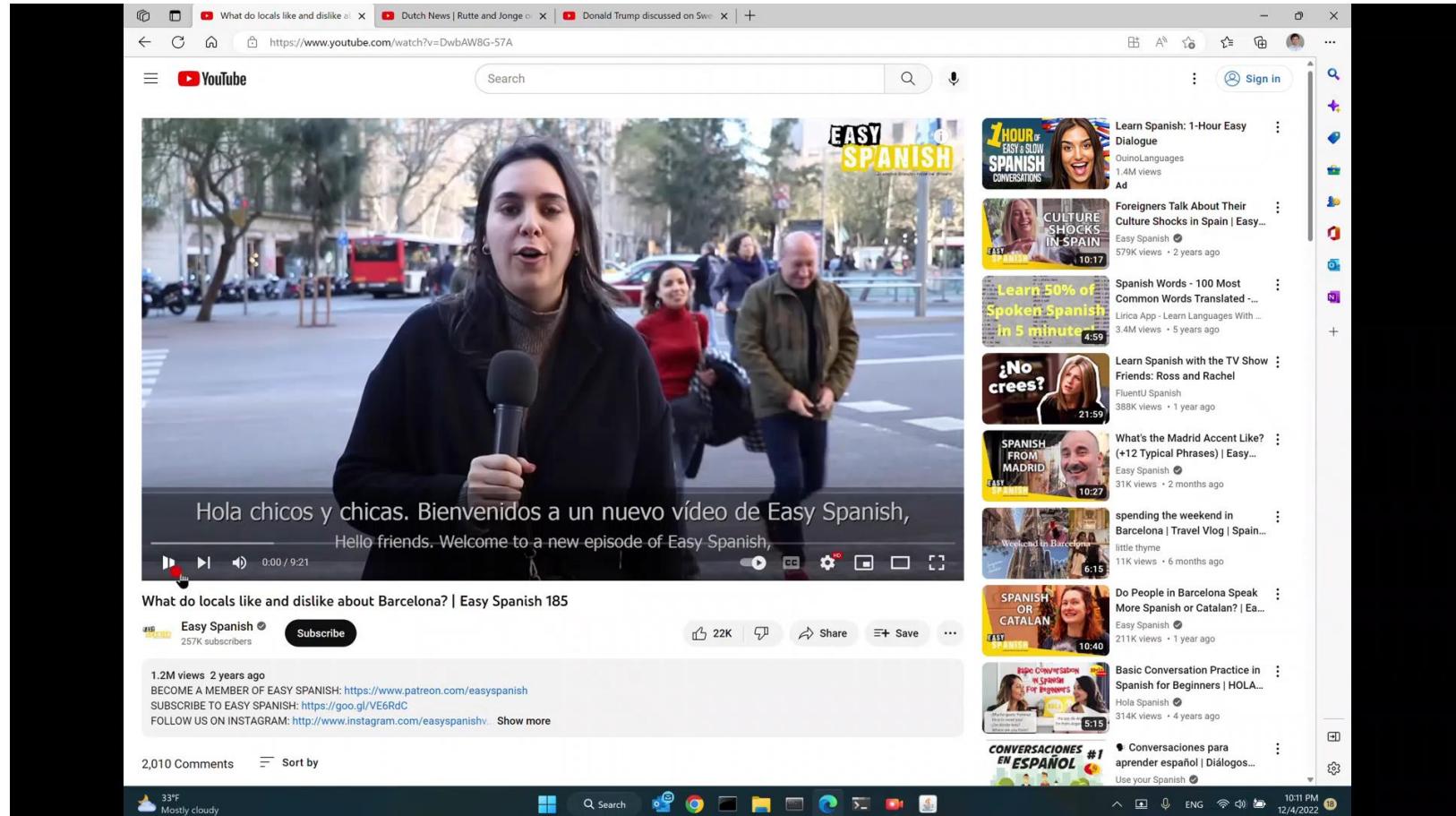
Language Expansion

- Every language has its own prediction and joint network, sharing the same encoder



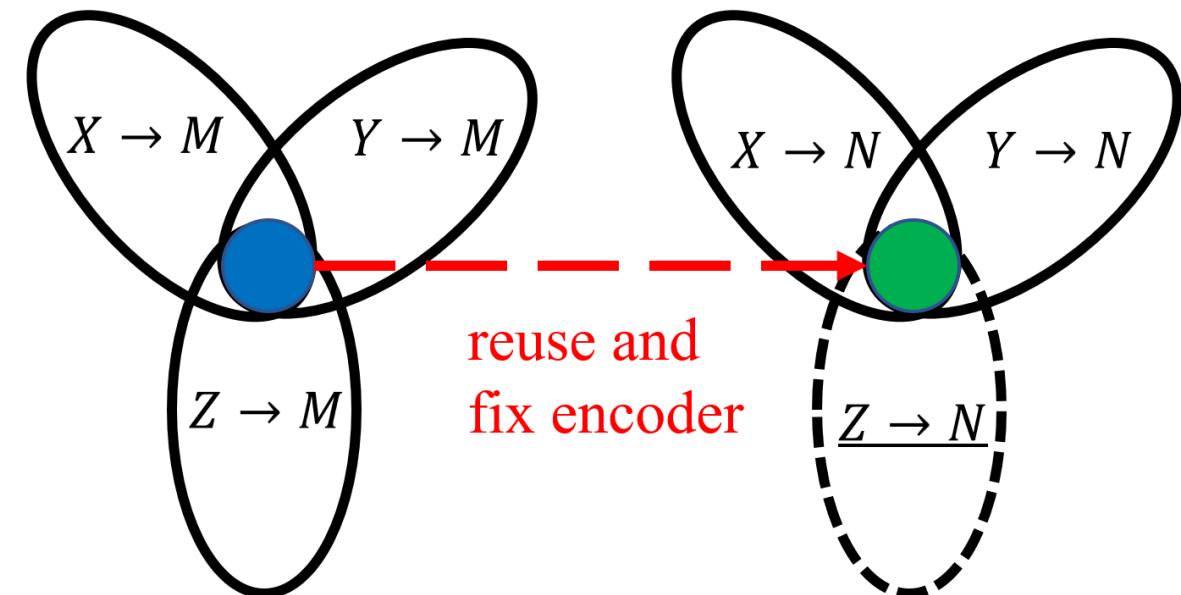
Zero-Shot Speech Translation

Trained only with English/German/Chinese->Chinese data, without observing any other language to Chinese.



Why Can SM^A2 Do the Zero-Shot Translation?

- The utterances in the interlingua space (circle) have the same semantic meaning.
- Encoder is frozen for a new language output.
- Utterances in the interlingua space learn to translate to the new target language even if the pair is not observed.
- Because of the calibration inside the language, the learning can be extended to other utterances in the unseen language (dashed area).



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.
 - Multilingual: configurable multilingual model
 - Leverage unpaired text: domain adaptation and speech/text joint training
 - Multi-talker ASR: (token-level) serialized output training
 - Speech translation: streaming multilingual speech model

Future Directions

- Self-supervised learning: leveraging unlabeled data.
- Multi-modality: e.g., VATLM: visual-audio-text joint training.
- High quality multilingual models
- E2E ASR with extended functions: multi-talker, multi-channel, and speaker diarization
- E2E with knowledge integration
- E2E for more downstream tasks



Thank You!