语音翻译：从前沿研究到产品创新

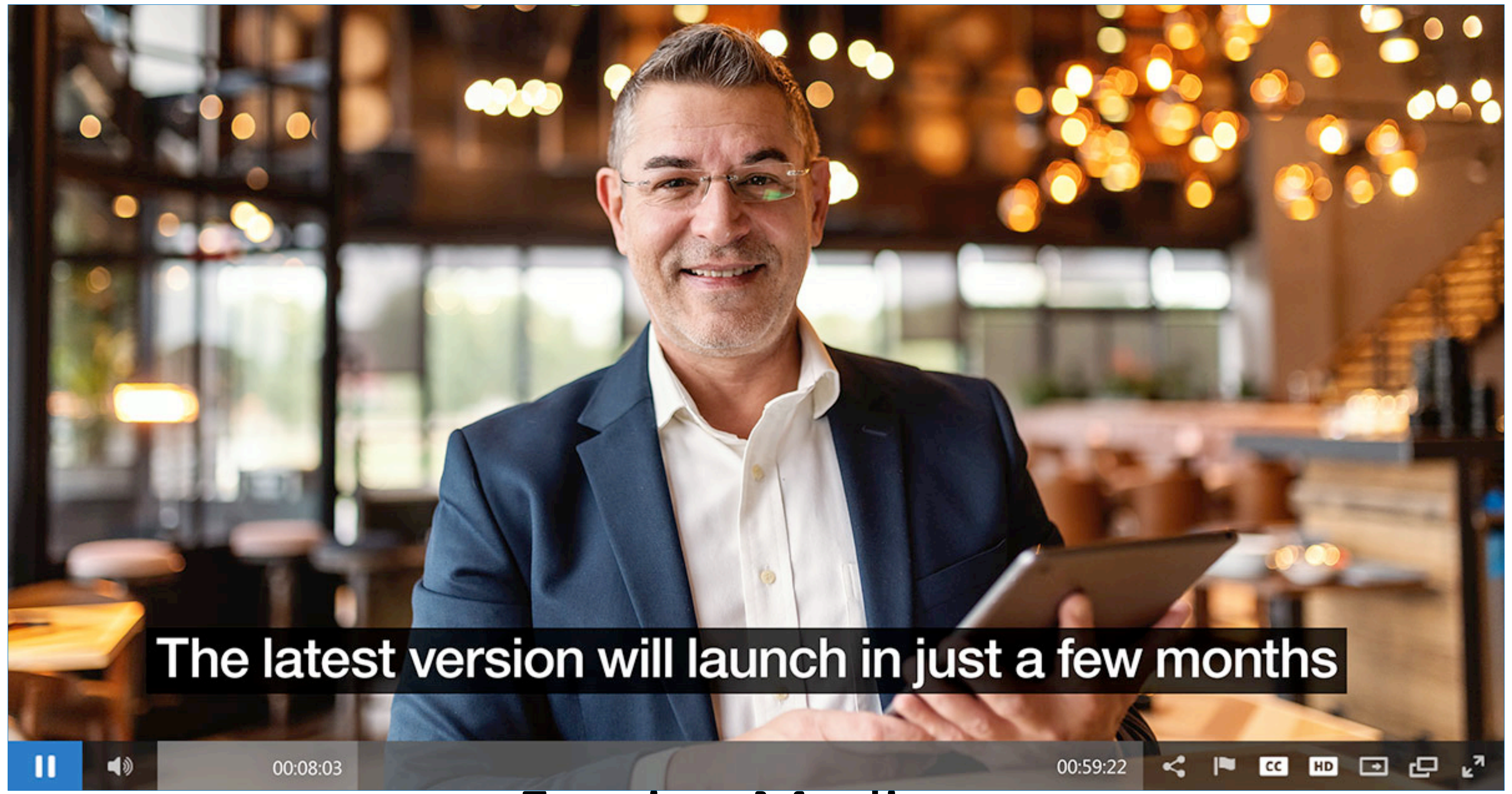
Speech Translation

李磊

字节跳动人工智能实验室

2021/6/6

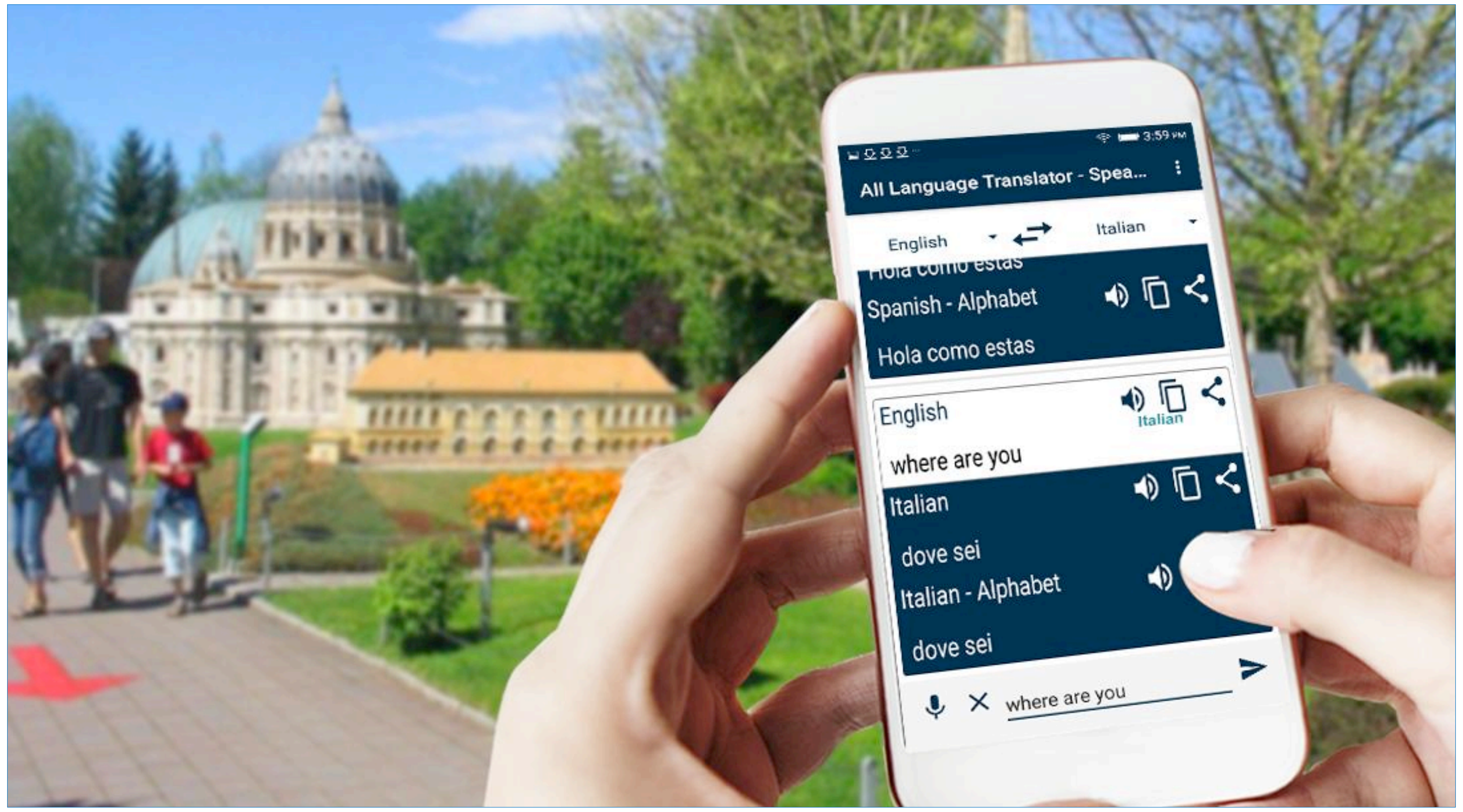
Cross Language Barrier with Machine Translation



The latest version will launch in just a few months

Foreign Media

Global Conferences



All Language Translator - Spea...

English ⇄ Italian

Hola como estas

Spanish - Alphabet

Hola como estas

English

where are you

Italian

dove sei

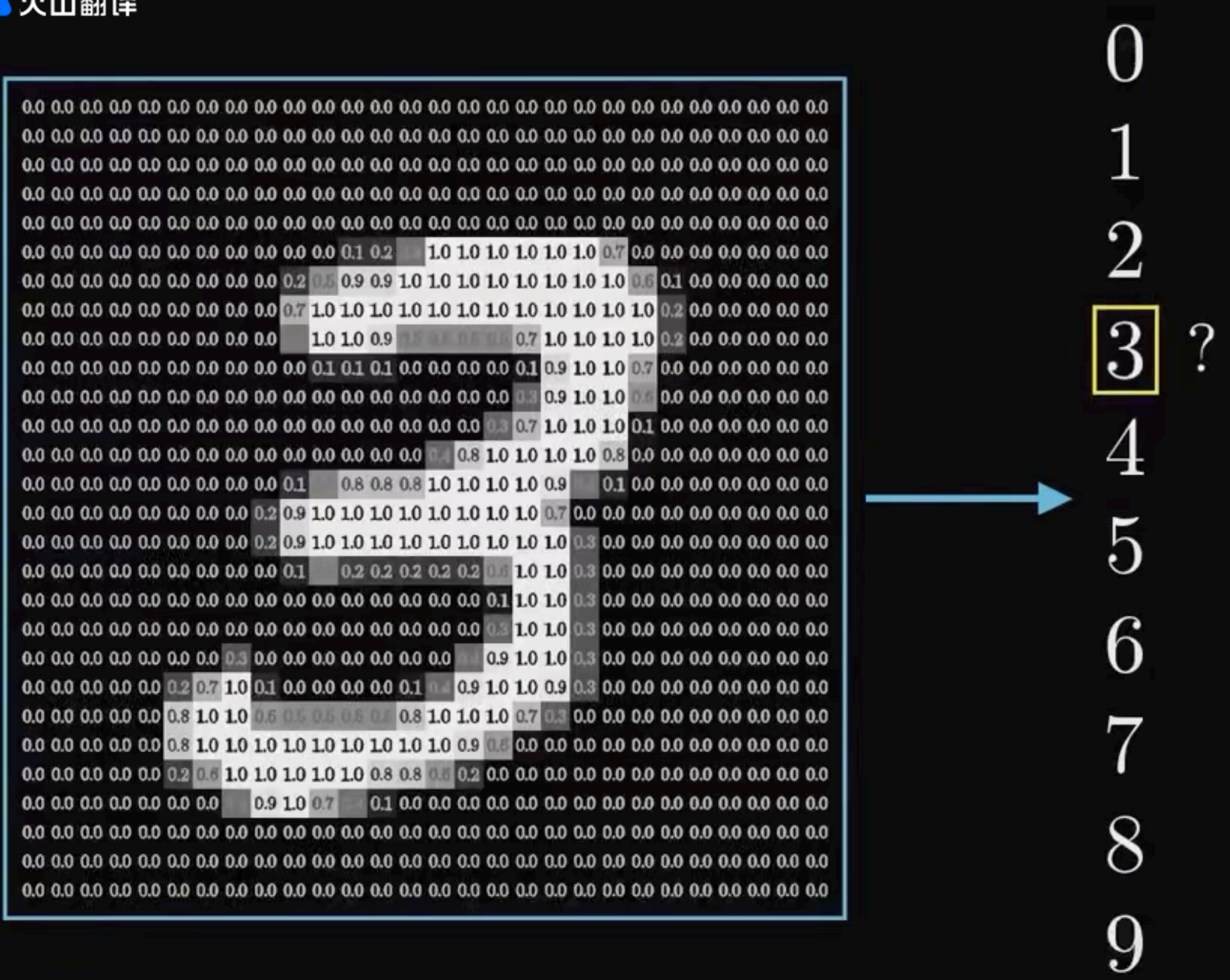
Italian - Alphabet

dove sei

where are you

Tourism

International Trade



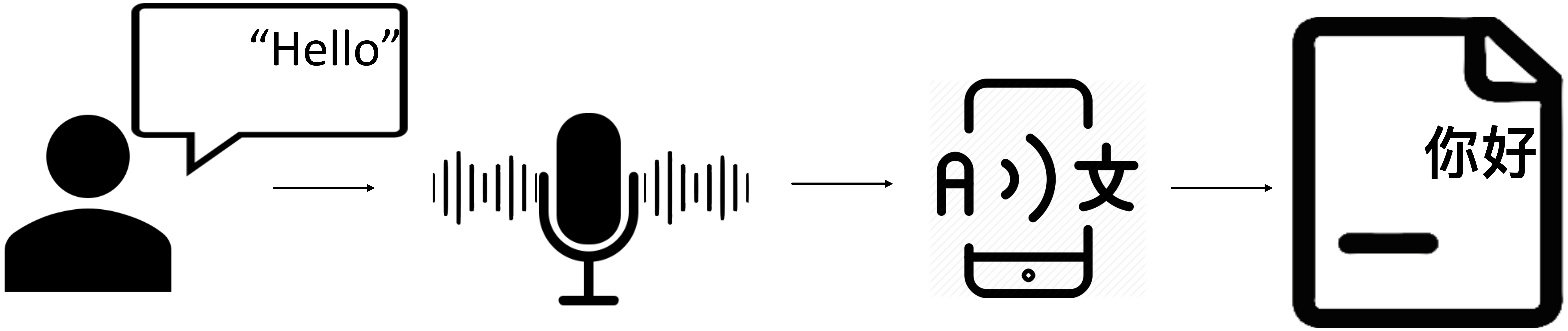
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	[illegible]	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6	0.1	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	[illegible]	1.0	1.0	0.9	0.5	0.5	0.5	0.6	0.7	1.0	1.0	1.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.9	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.9	1.0	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.7	1.0	1.0	1.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.8	1.0	1.0	1.0	1.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	[illegible]	0.8	0.8	0.8	1.0	1.0	1.0	1.0	0.9	[illegible]	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	[illegible]	0.2	0.2	0.2	0.2	0.2	0.6	1.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	[illegible]	0.9	1.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.0	0.1	0.0	0.0	0.0	0.0	0.1	0.4	0.9	1.0	1.0	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.8	1.0	1.0	0.6	0.5	0.5	0.5	0.5	0.8	1.0	1.0	1.0	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.2	0.6	1.0	1.0	1.0	1.0	1.0	0.8	0.8	0.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	[illegible]	0.9	1.0	0.7	[illegible]	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Outline

1. Overview: ST Problem and Challenge
2. What is a better model for ST?
3. Better training strategy for ST?
4. New ST-powered Products

Speech-to-Text Translation(ST)

- source language ***speech(audio)*** → target lang ***text***



Application Type

- (Non-streaming) ST
  非流式语音翻译
- Streaming ST
  流式语音翻译

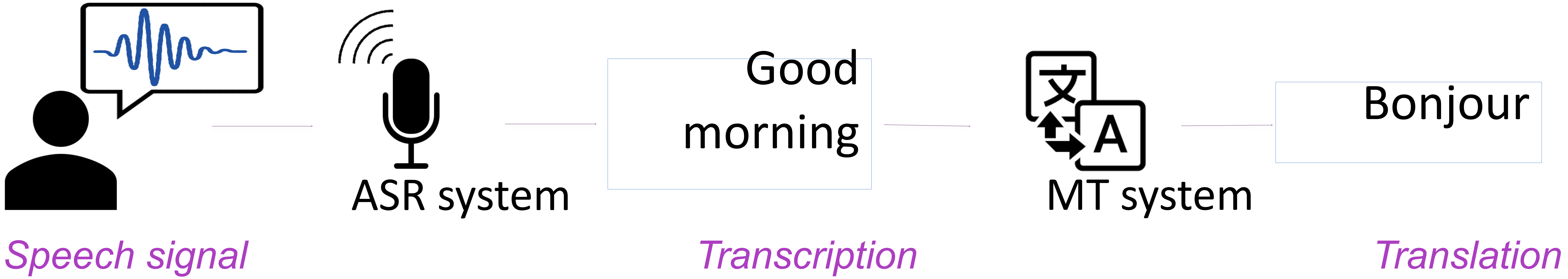
System

- Cascaded ST
  级联语音翻译
- End-to-end ST
  端到端语音翻译

Cascaded ST System

- Challenges:

1. **Computationally inefficient**
2. **Error propagation**: Wrong transcription ⍰ Wrong translation



Speech signal

ASR system

Good morning

Transcription

MT system

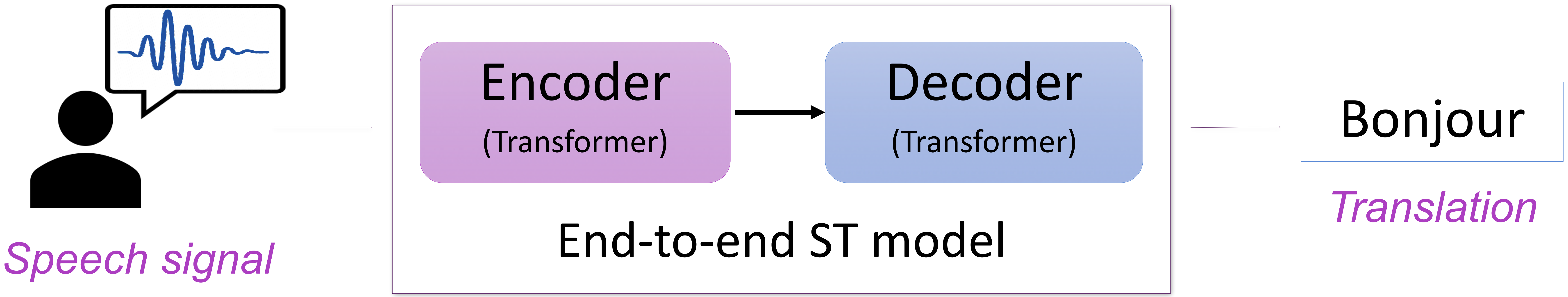
Bonjour

Translation

do at this *and see if it works for you* ⍰ **这样做**，看看它是否对你有用

duet this *and see if it works for you* ⍰ **二重奏一下**，看看它是否对你有用

End-to-end ST Model

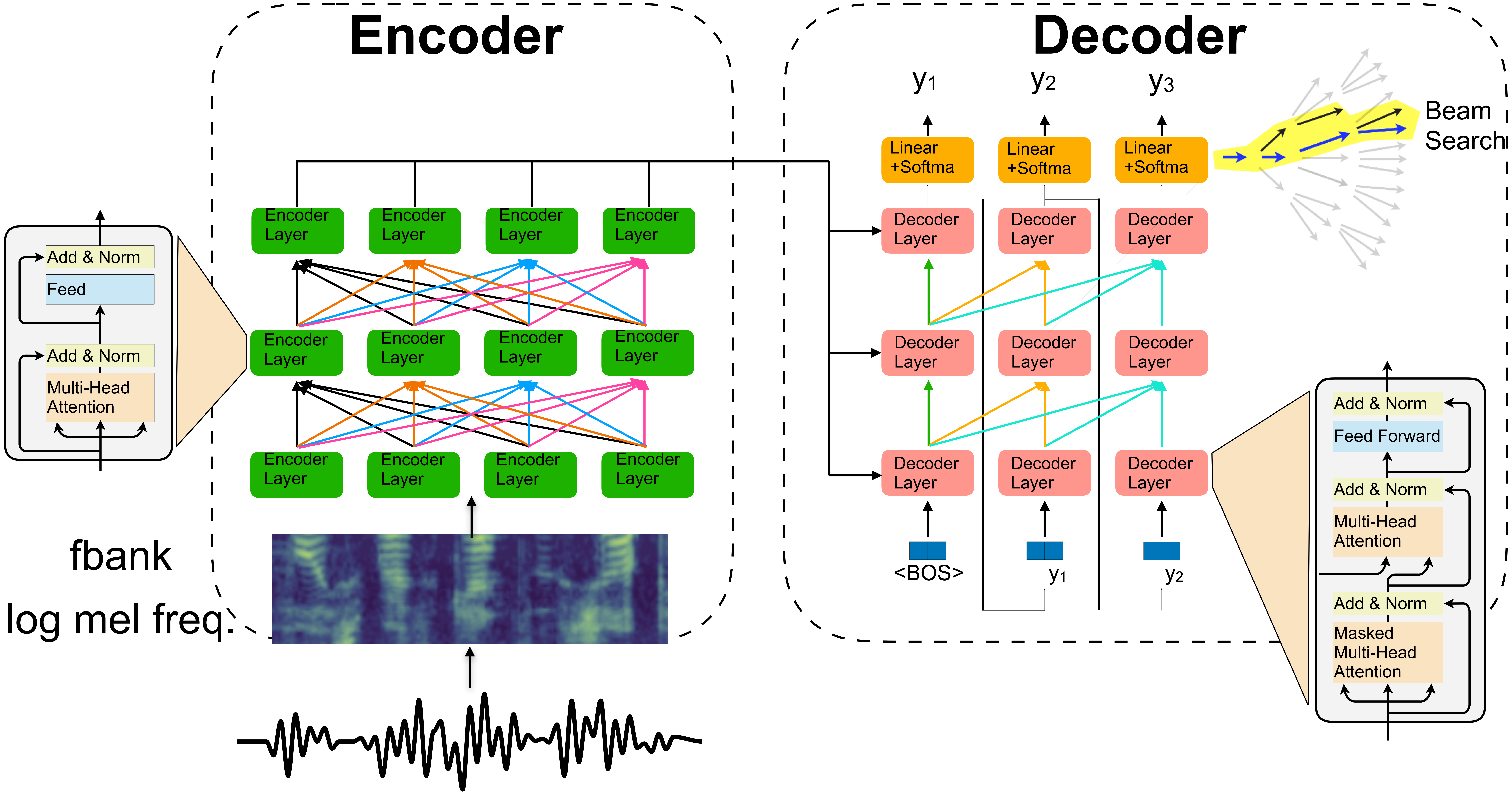


- Single model to produce text translation from speech
- Popular model: Encoder-Decoder architecture (e.g. Transformer)
- Advantage:
  - Reduced latency, simpler deployment
  - Avoid error propagation

[1] Bérard et al., Listen and translate: A proof of concept for end-to-end speech-to-text translation. 2016

Basic Speech Translation Architecture (Same as MT)

Transformer-based: N-layer encoder, M-layer decoder



Challenge

- Data scarcity - lack of large parallel corpus
- Modality disparity between audio and text
- Require low latency for product serving

Approaches for End-to-end ST

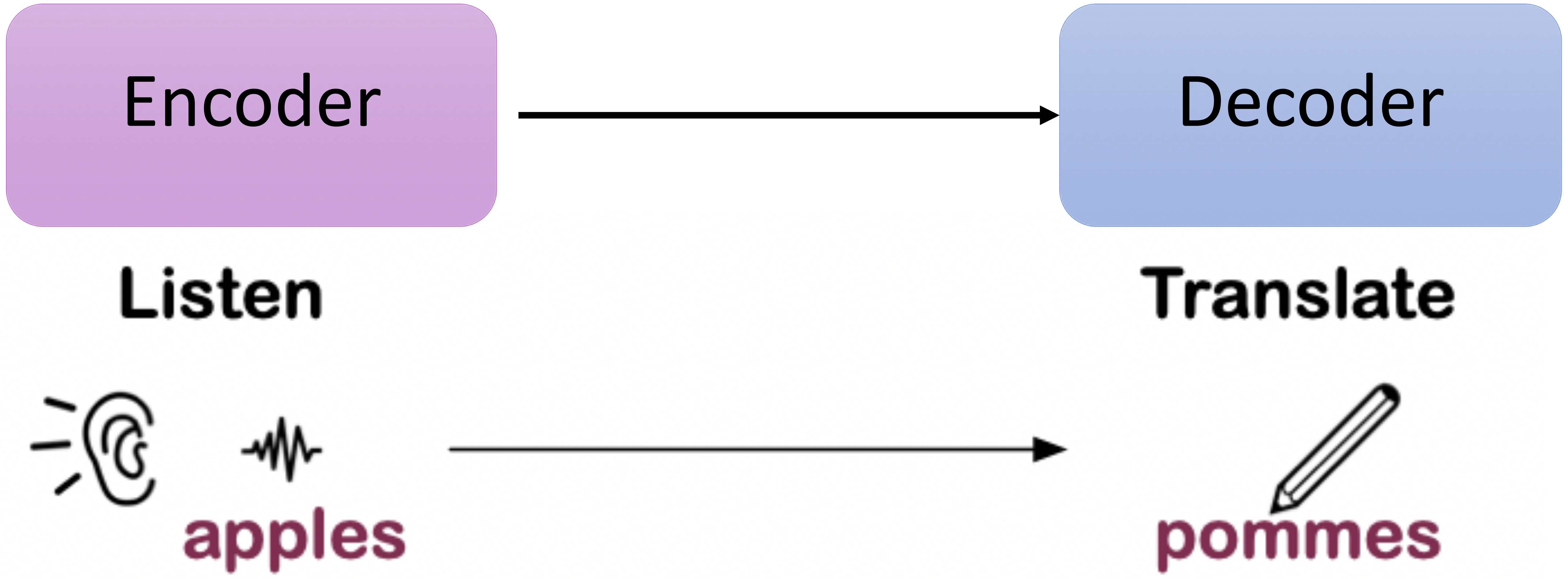
- Model
  - Better Encoder: LUT [AAAI 2021a] Chimera[ACL 2021a]
  - Better Decoder: COSTT[AAAI 2021b]
- Training technique
  - Audio pre-training: Wave2Vec2.0[Baevski et al 2021]
  - Progressive multi-task training: XSTNet [Interspeech 2021]
- Speed-up Inference (not in this talk)
  - Parallel Decoding: GLAT [ACL 2021b]
  - GPU optimization: LightSeq [NAACL2021]

LUT

Listen, Understand and Translate: Triple Supervision Decouples End-to-end Speech-to-text Translation

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei Li

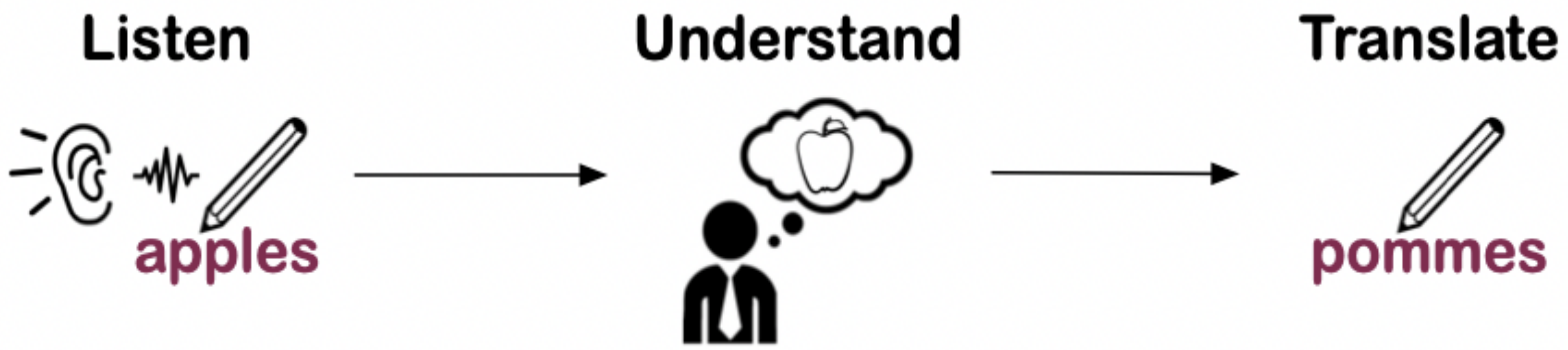
Drawbacks of the Encoder-Decoder Structure



1. A single encoder is hard to capture the representation of audio for the translation.
2. Limited in utilizing the information of *“transcription”* in the training.

Motivation: Mimic human's behavior

Question: How human translate?

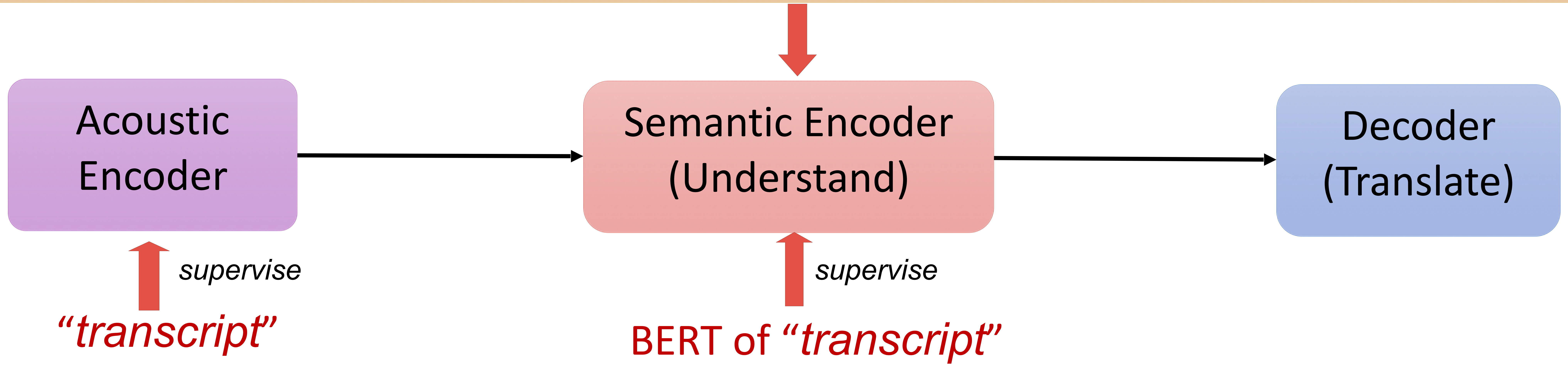


"Listen-Understand-Translate"(LUT) model based motivated by human's behavior

Motivation of Better Encoding

Drawback 1: A single encoder is not enough.

Idea 1: Introduce a semantic encoder



Acoustic Encoder

Semantic Encoder (Understand)

Decoder (Translate)

supervise

"transcript"

supervise

BERT of *"transcript"*

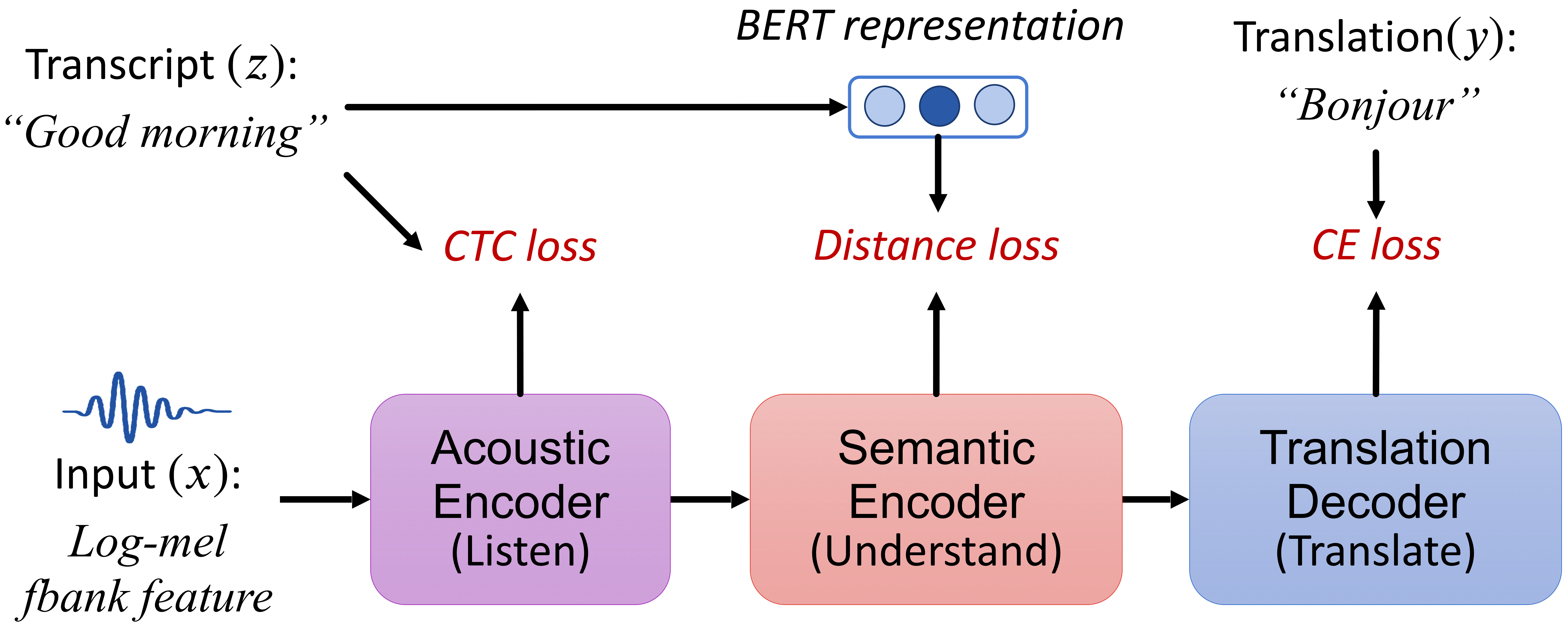
Drawback 2: Limit in using "transcript" info.

Idea 2: Utilizing the pre-trained representation (e.g. BERT) of the "transcript" to learn the semantic feature.

LUT for End-to-end ST

Training data: triples of

<speech, transcript_text, translate_text>

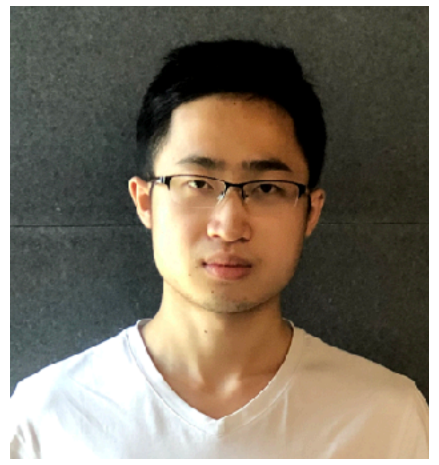


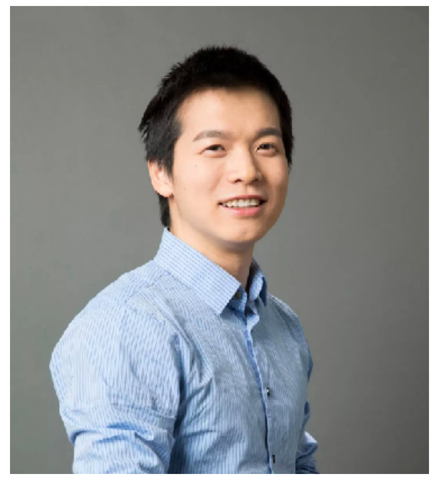
Listen, Understand and Translate [Q. Dong, R. Ye, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

Chimera

Learning Shared Semantic Space for Speech-to-Text Translation

Chi Han, Mingxuan Wang, Heng Ji, Lei Li

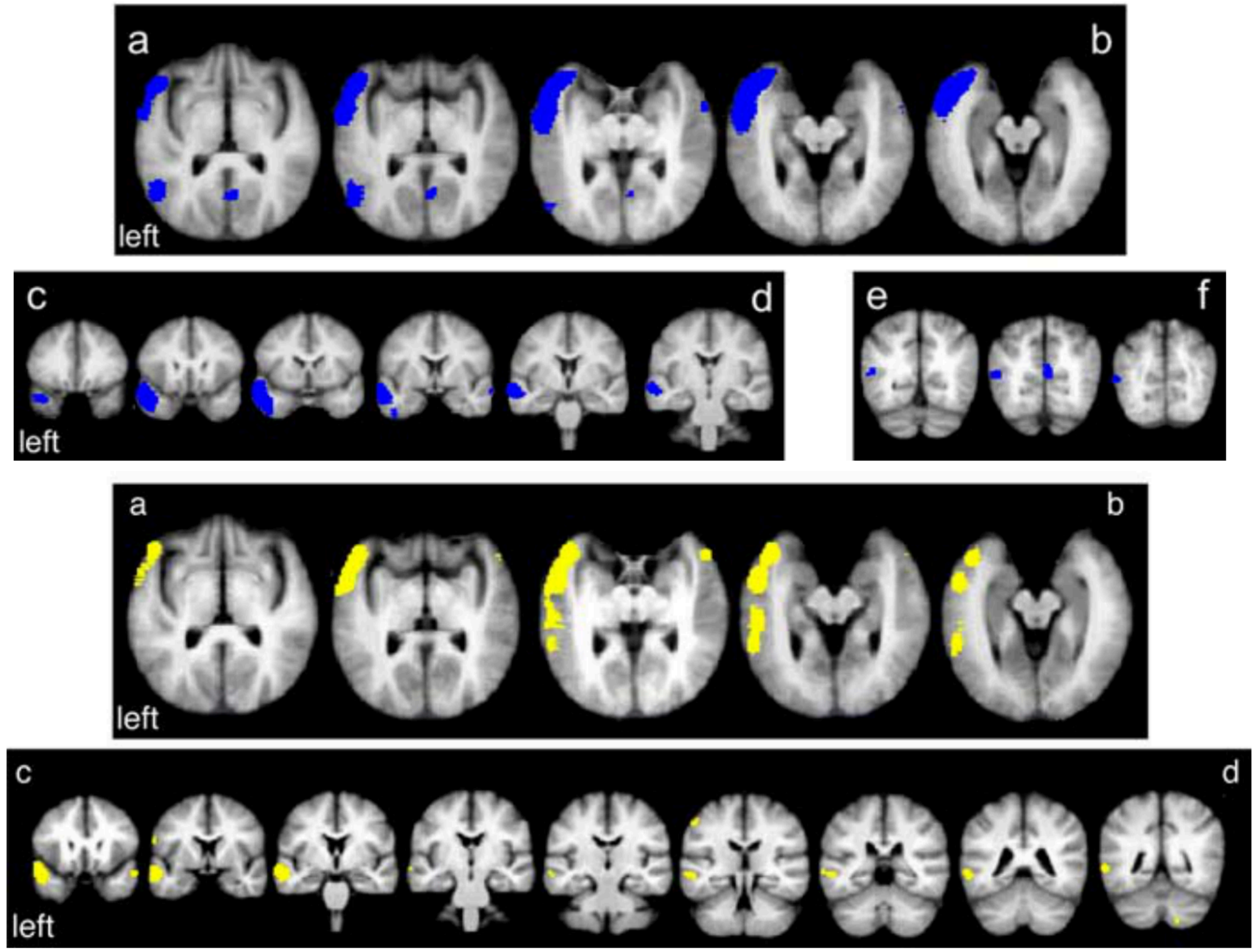




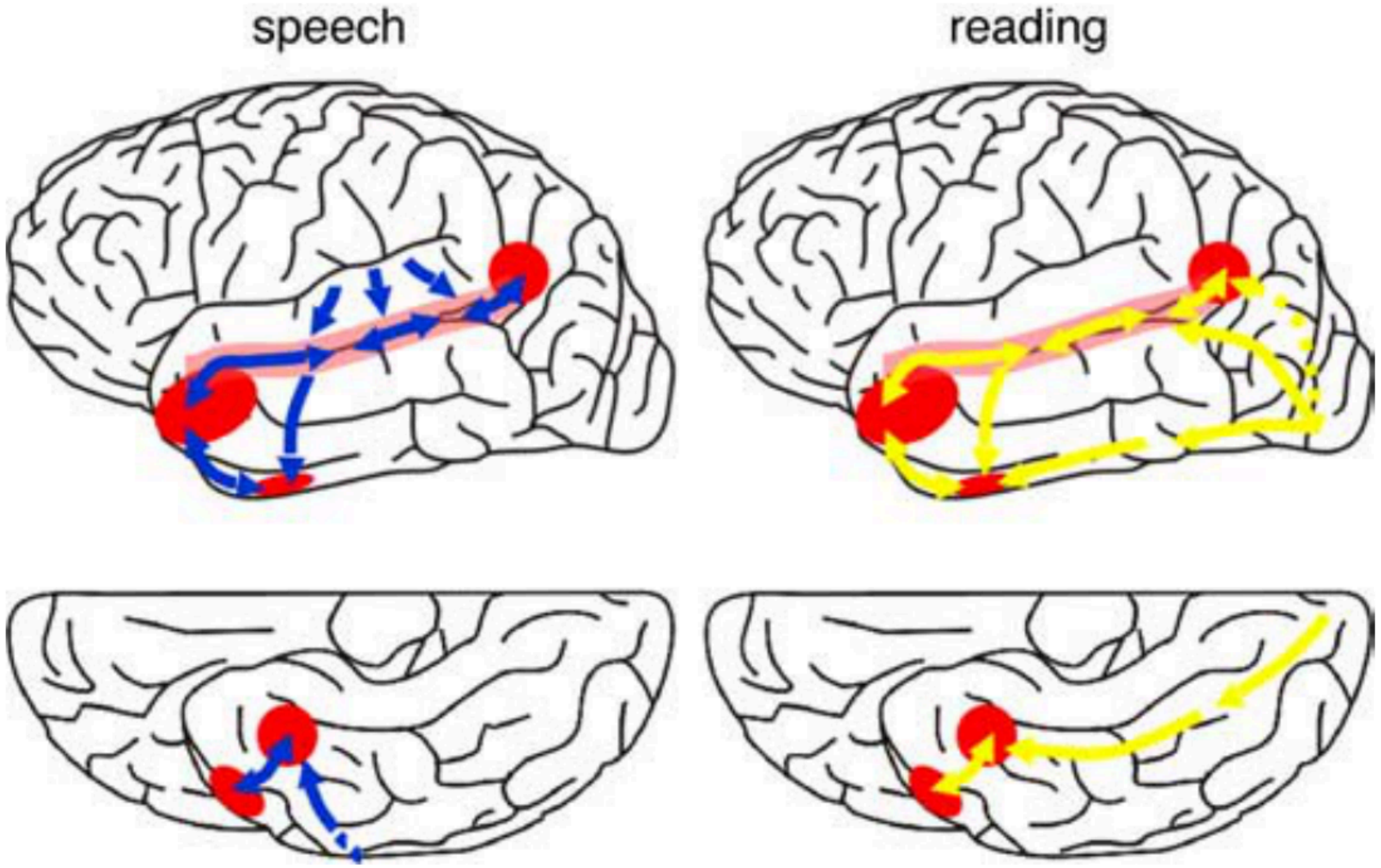
Paper: https://arxiv.org/abs/2105.03095
Code: https://github.com/Glaciohound/Chimera-ST

Insights from Cognitive Neuroscience

Speech and text interfere with each other in brain[1]



activation map



processing paths

<u>Convergence sites</u> of *speech* (blue) and *text* (yellow)

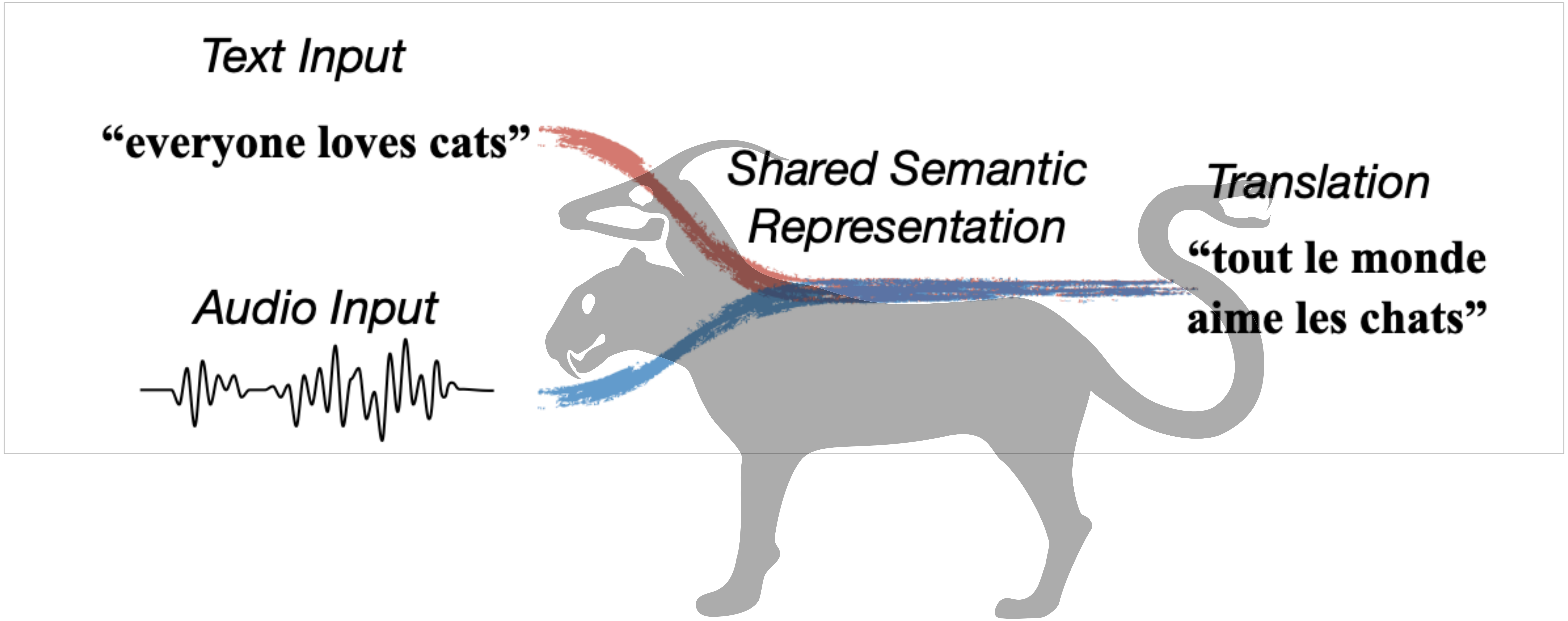
[1] Van Atteveldt, Nienke, et al. "Integration of letters and speech sounds in the human brain." *Neuron* 43.2 (2004): 271-282.

[2] Spitsyna, Galina, et al. "Converging language streams in the human temporal lobe." *Journal of Neuroscience* 26.28 (2006): 7328-7336.

Idea: Bridging the Speech-Text modality gap with Shared Semantic Representation

ST triple data:

<speech, transcript_text, translate_text>

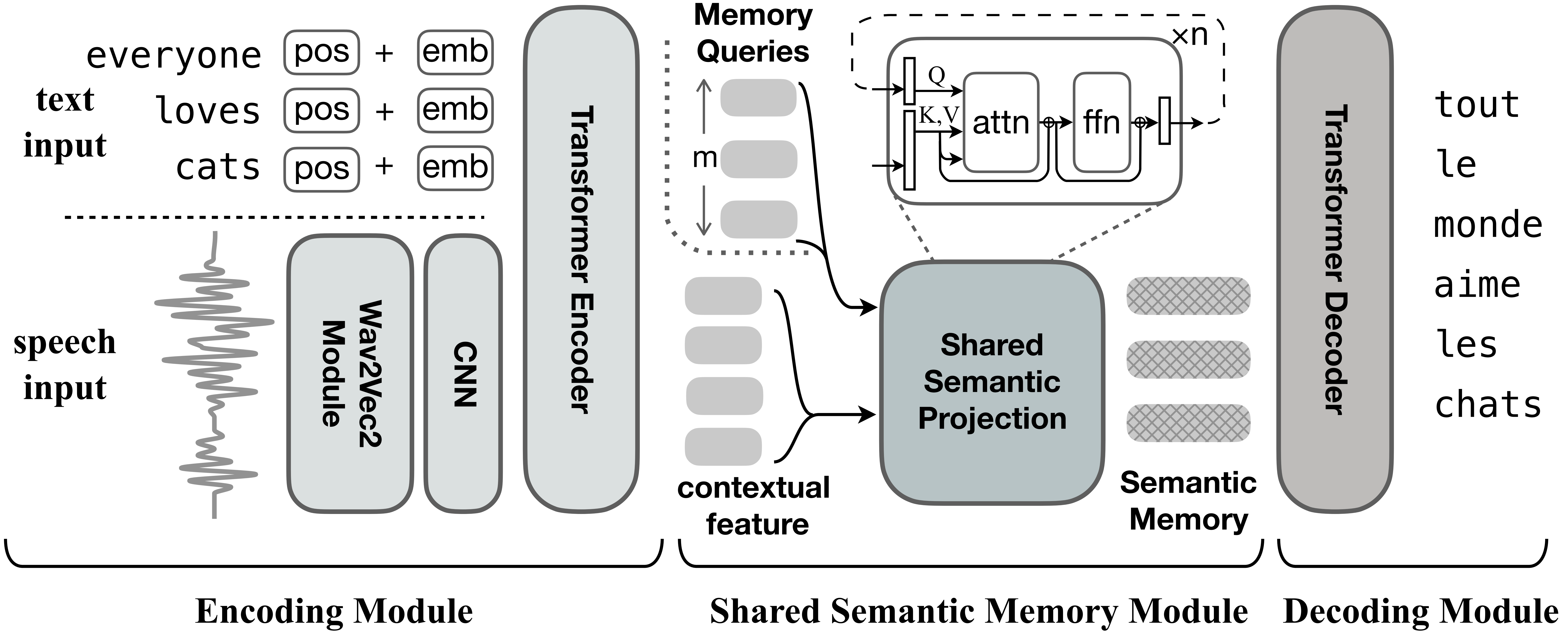

Text Input
"everyone loves cats"
Audio Input
Shared Semantic Representation
Translation
"tout le monde aime les chats"

Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Chimera Model for ST

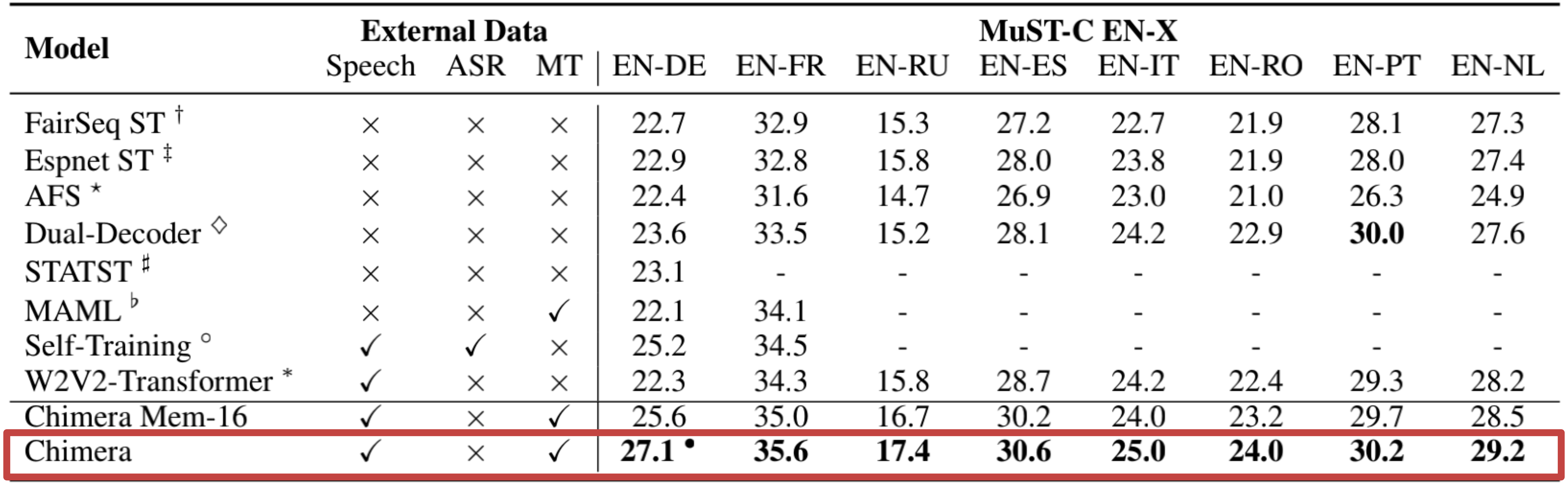
Training with auxiliary objectives: ST + MT + Contrastive loss

Benefit: able to exploit large external MT data



Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Chimera achieves the best (so far) BLEU on all languages in MuST-C



Model	External Data			MuST-C EN-X							
	Speech	ASR	MT	EN-DE	EN-FR	EN-RU	EN-ES	EN-IT	EN-RO	EN-PT	EN-NL
FairSeq ST †	×	×	×	22.7	32.9	15.3	27.2	22.7	21.9	28.1	27.3
Espnet ST ‡	×	×	×	22.9	32.8	15.8	28.0	23.8	21.9	28.0	27.4
AFS ⋆	×	×	×	22.4	31.6	14.7	26.9	23.0	21.0	26.3	24.9
Dual-Decoder ◇	×	×	×	23.6	33.5	15.2	28.1	24.2	22.9	**30.0**	27.6
STATST ♯	×	×	×	23.1	-	-	-	-	-	-	-
MAML ♭	×	×	✓	22.1	34.1	-	-	-	-	-	-
Self-Training ∘	✓	✓	×	25.2	34.5	-	-	-	-	-	-
W2V2-Transformer *	✓	×	×	22.3	34.3	15.8	28.7	24.2	22.4	29.3	28.2
Chimera Mem-16	✓	×	✓	25.6	35.0	16.7	30.2	24.0	23.2	29.7	28.5
Chimera	✓	×	✓	**27.1** •	**35.6**	**17.4**	**30.6**	**25.0**	**24.0**	**30.2**	**29.2**

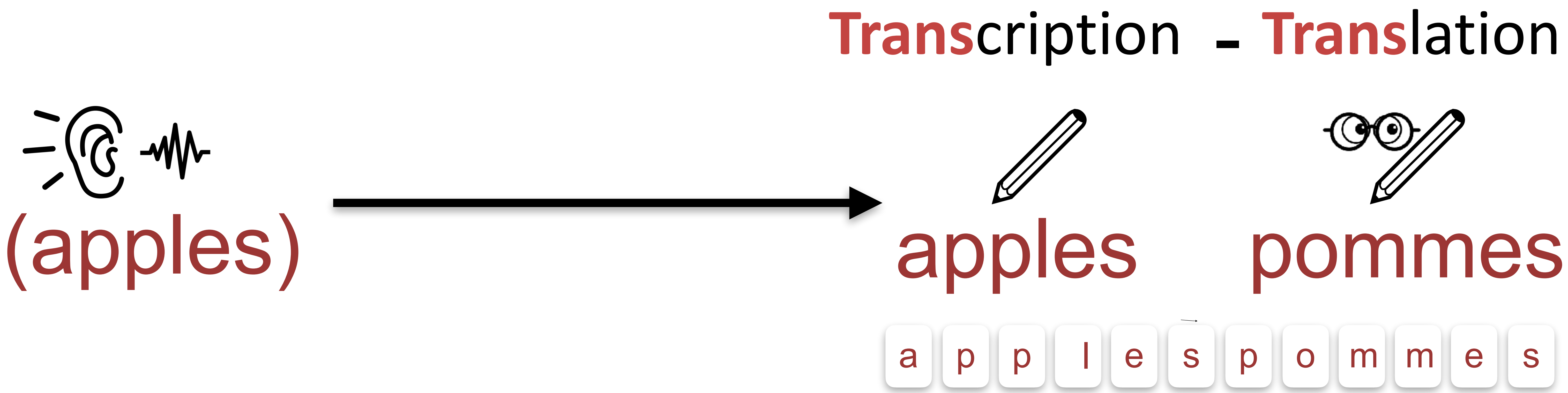
Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Consecutive Decoding for Speech-to-text Translation

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei Li

Goal: Seamless Trans-trans🤗

Question: How to help the model take notes like human interpreter?



(apples)

Transcription - Translation

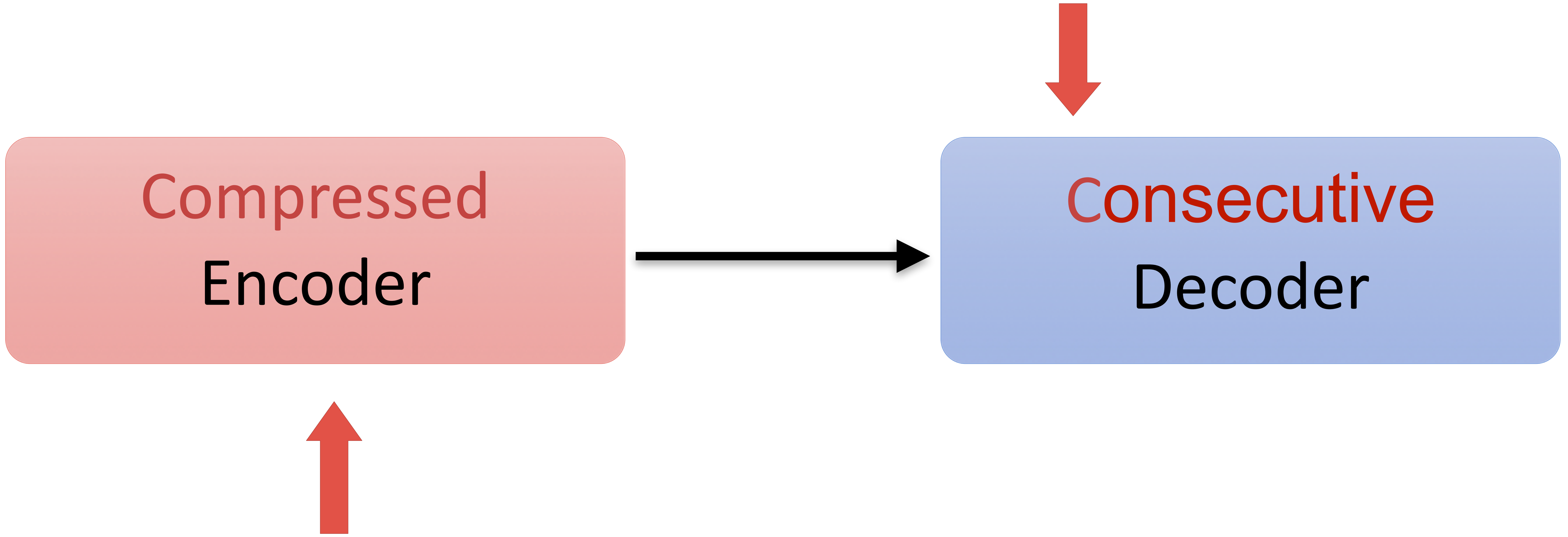
apples pommes

a p p l e s p o m m e s

We design "COnSecutive Transcription and Translation"(COSTT) based on interpreter's noting behavior to help the model memory.

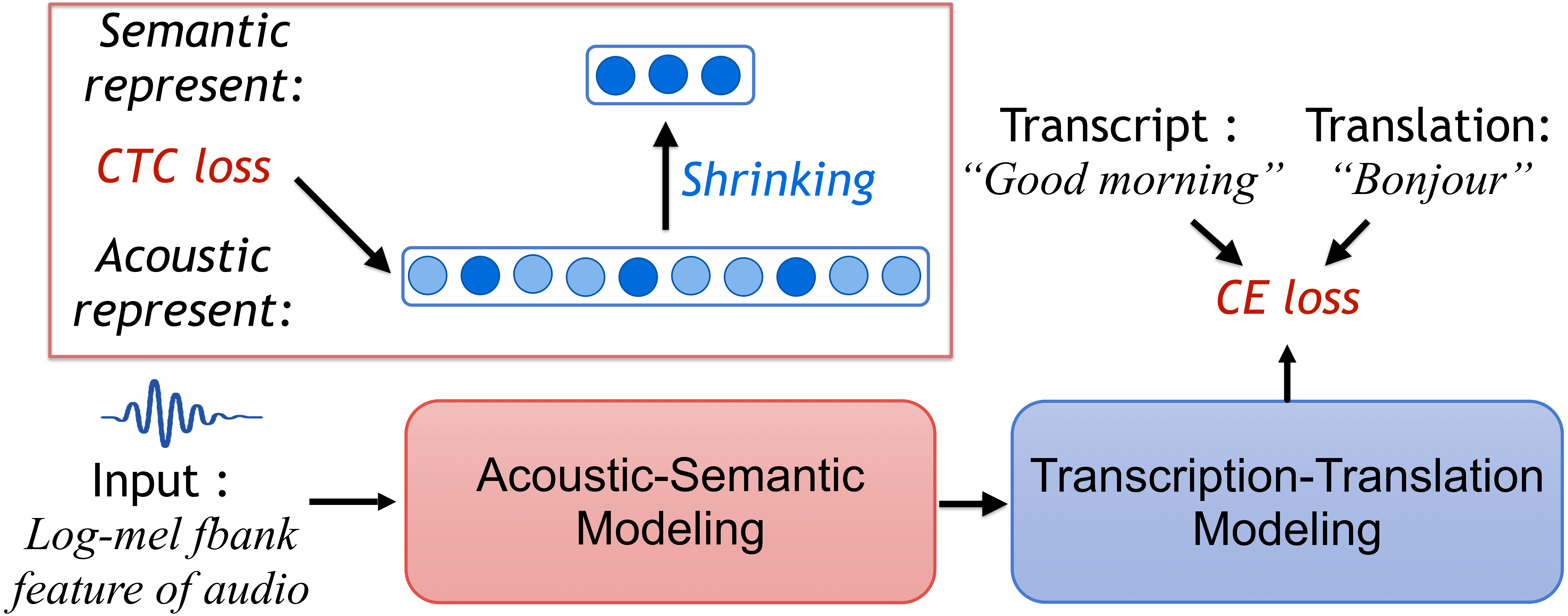
Motivation of Better Decoding

Problem1: How to give the decoder hints?
Idea 1: Introduce a consecutive decoder for trans-trans.



Problem2: Long acoustic sequence is challenging for the encoder!
Idea 2: Introduce a compressed encoder to relief the model memory.

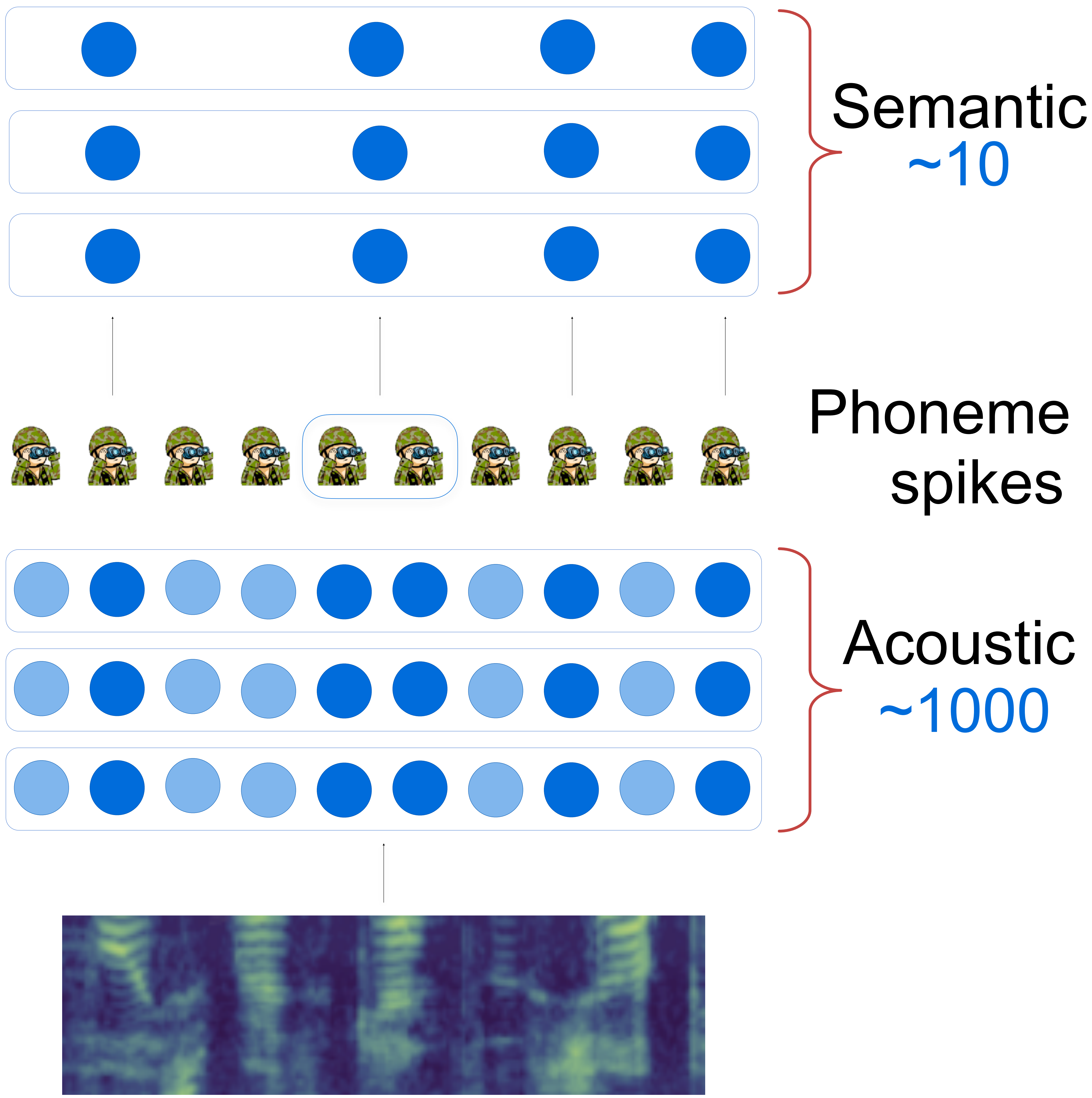
COSTT for ST



Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

Advantages of COSTT

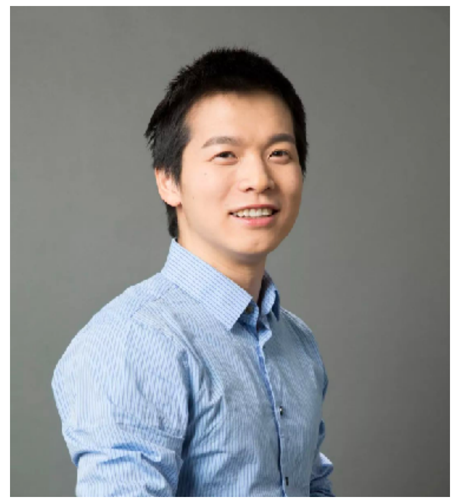
- Unified training with both transcript and translation text
- Reduced encoder output size with CTC-guided shrinking
- Able to pre-train the decoder with external MT parallel data



Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

End-to-end Speech Translation via Cross-modal Progressive Training

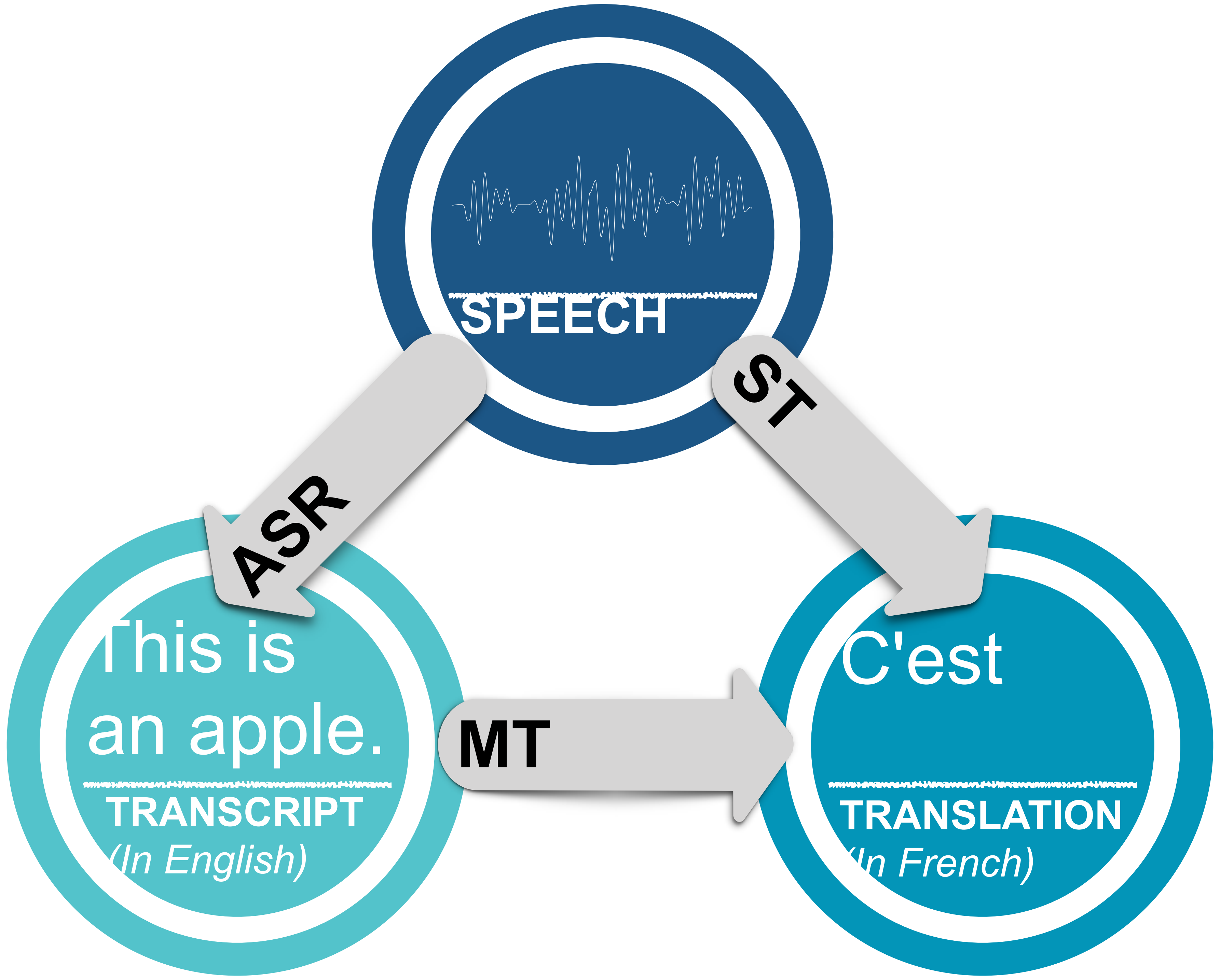
Rong Ye, Mingxuan Wang, Lei Li



- Link: https://arxiv.org/abs/2104.10380

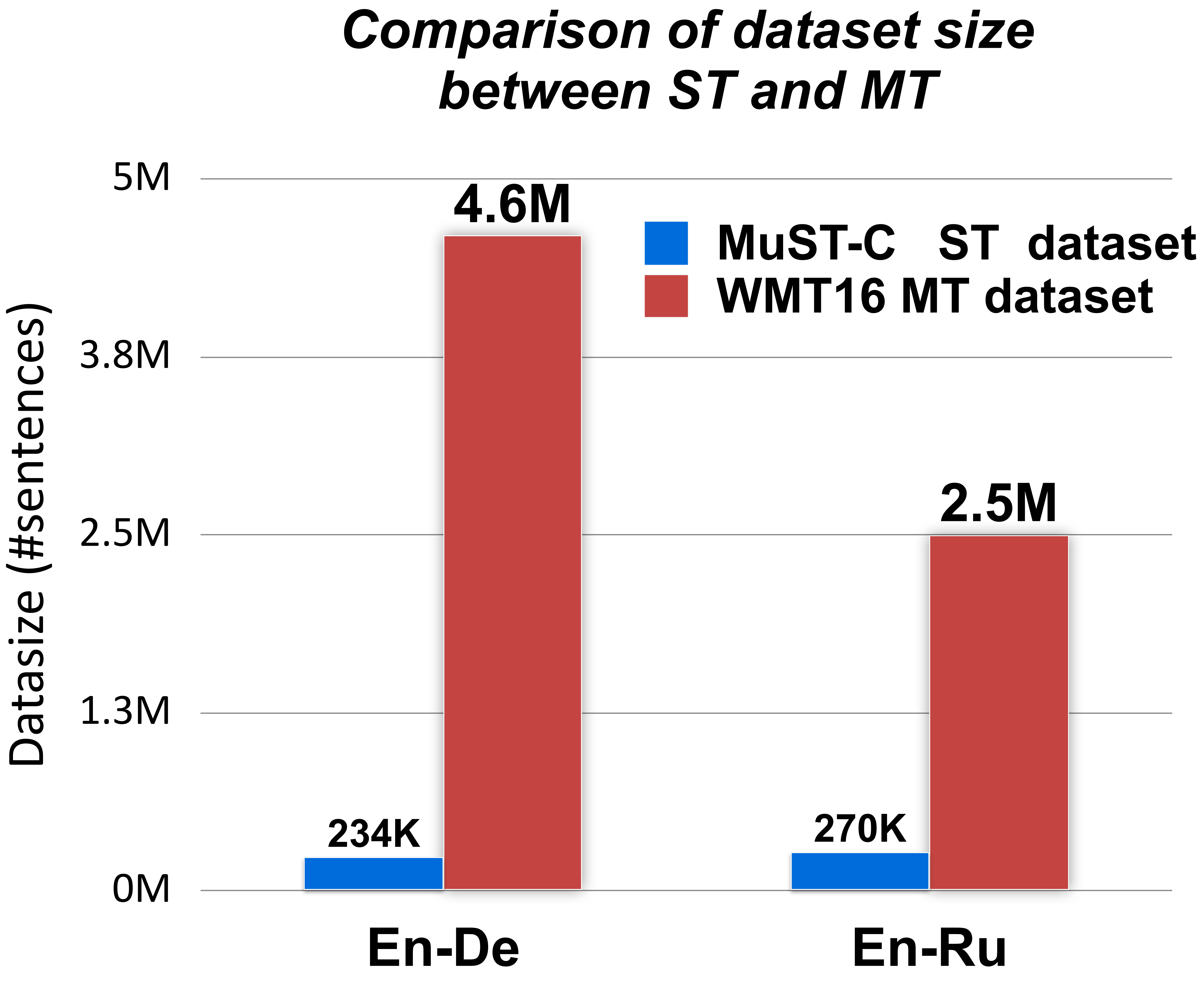
Idea 1: Multi-task Training

Goal: To fully utilize the existing *<Speech, Transcript, Translation>* supervision.



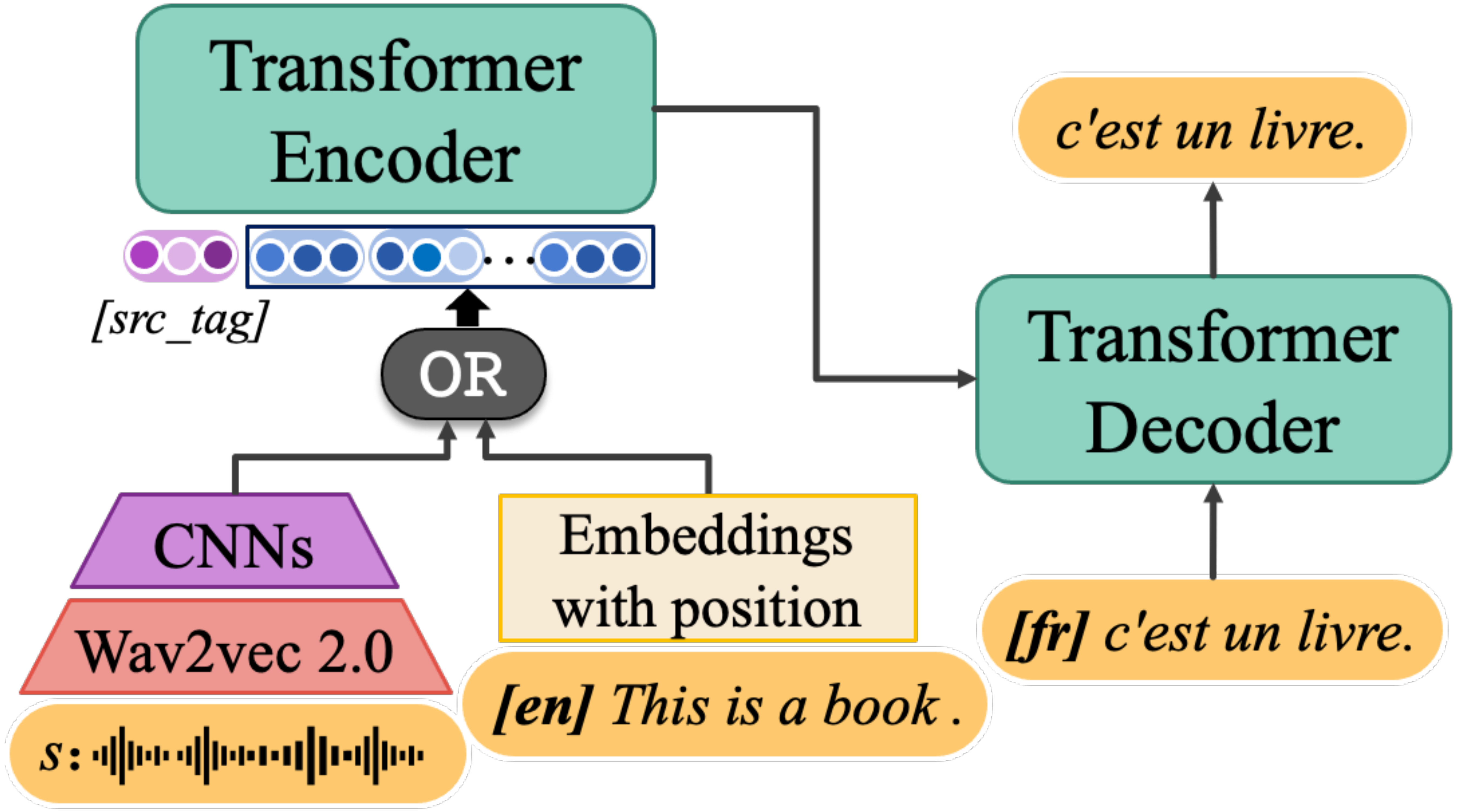
Decomposed into three sub-tasks with parallel supervision, ST, ASR and MT.

Idea 2: Using large-scale MT data



🤔 How to introduce MT data *with much larger scale* to improve ST performance?

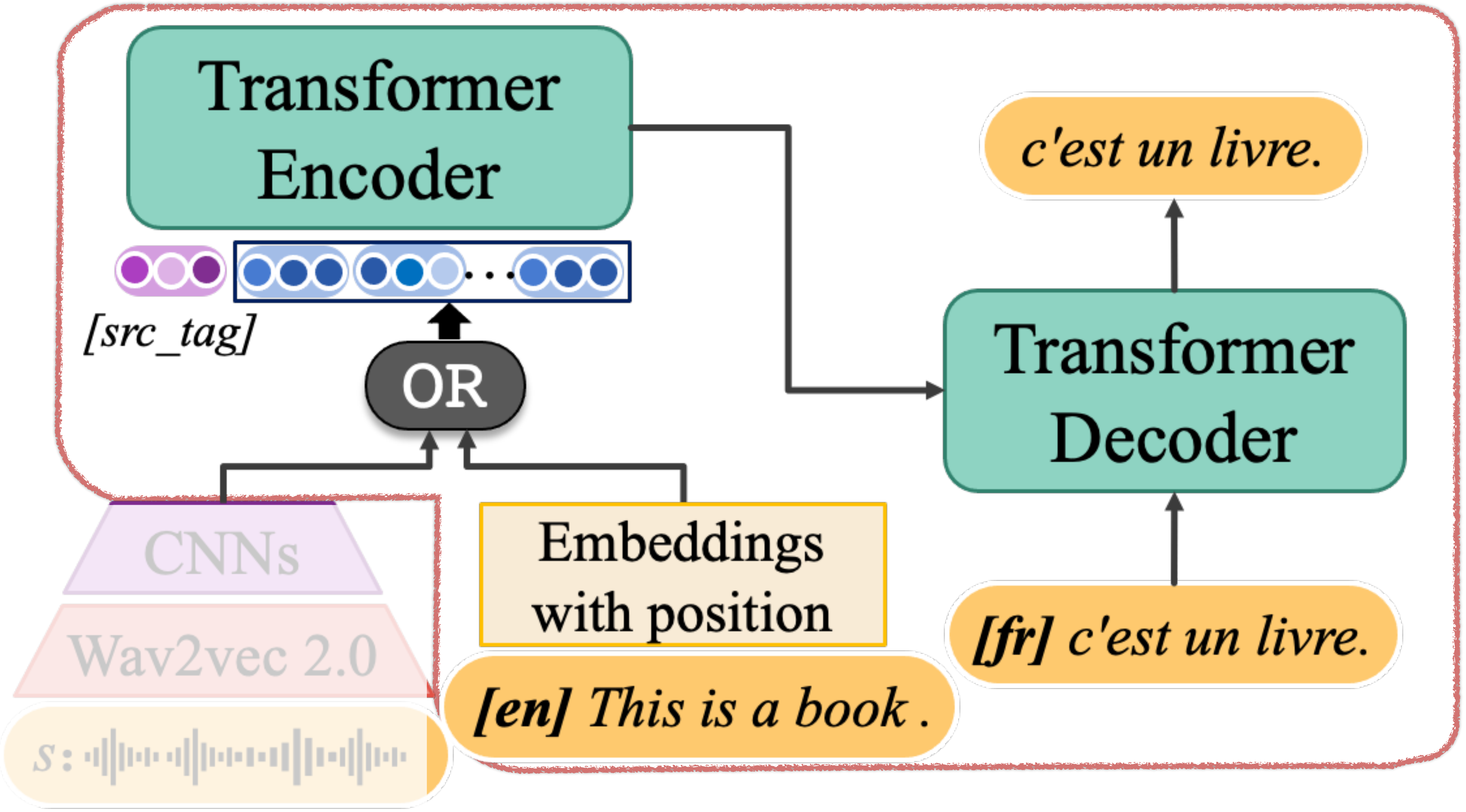
Cross Speech-Text Network (XSTNet)



End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Supports to train MT data

- ☑ Transformer MT model
- ☑ We can add **more external MT data** to train Transformer encoder & decoder



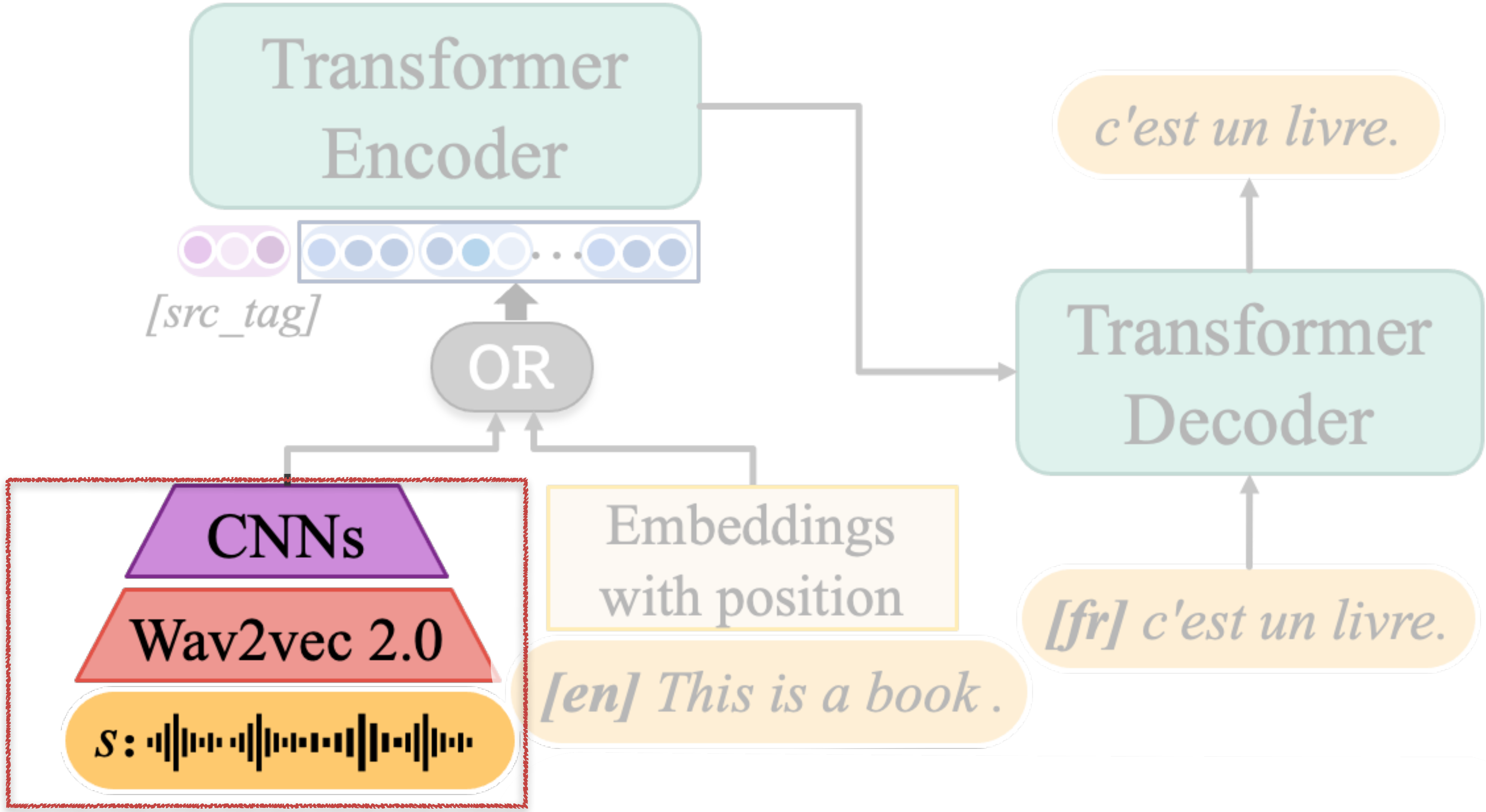
End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Supports inputs of two modalities

☑ Wav2vec2.0[1] as the acoustic encoder

☑ We add two convolution layers with 2-stride to shrink the length.

≈Acoustic encoder



[1] wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020

Language indicator strategy

- We use language indicators to distinguish different tasks.

Tasks	Source input	Target output
MT	**<en>** This is a book.	**<fr>** c'est un livre.
ASR	**<audio>**	**<en>** This is a book.
ST	**<audio>**	**<fr>** c'est un livre.

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Progressive Multi-task Training

Large-scale MT pre-training

Using **external MT** D_{MT-ext}

↓

Multi-task Finetune

Using **(1) external MT** D_{MT-ext}

(2) D_{ST} with *<speech, translation>*

(3) D_{ASR} with *<speech, transcript>*

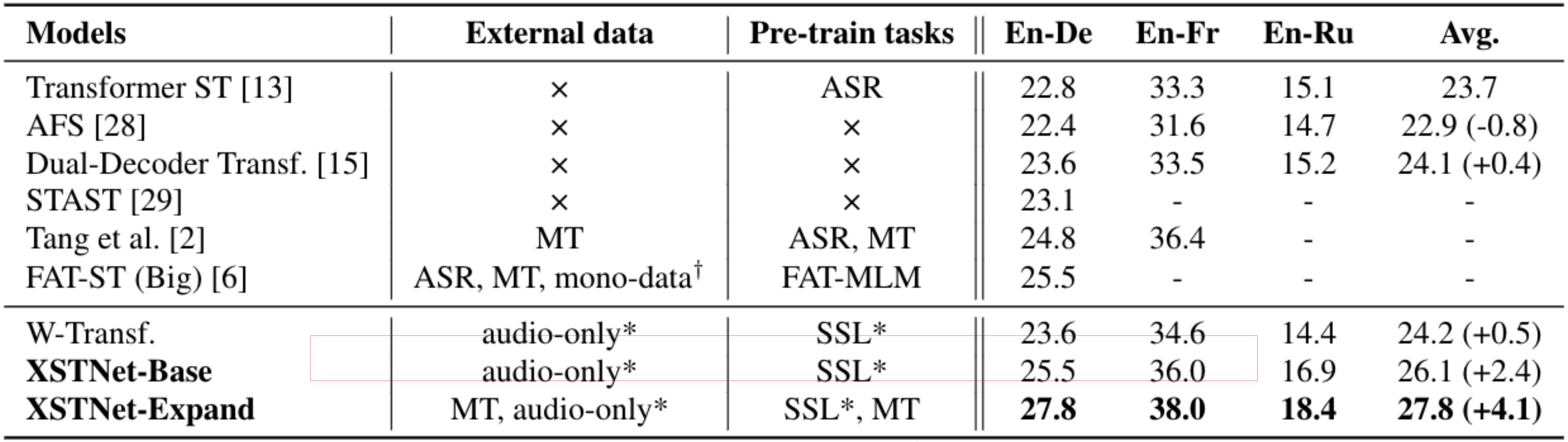
Progressive:

Don't stop

training D_{MT-ext}

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

XSTNet achieves State-of-the-art Performance



Models	External data	Pre-train tasks	En-De	En-Fr	En-Ru	Avg.
Transformer ST [13]	×	ASR	22.8	33.3	15.1	23.7
AFS [28]	×	×	22.4	31.6	14.7	22.9 (-0.8)
Dual-Decoder Transf. [15]	×	×	23.6	33.5	15.2	24.1 (+0.4)
STAST [29]	×	×	23.1	-	-	-
Tang et al. [2]	MT	ASR, MT	24.8	36.4	-	-
FAT-ST (Big) [6]	ASR, MT, mono-data†	FAT-MLM	25.5	-	-	-
W-Transf.	audio-only*	SSL*	23.6	34.6	14.4	24.2 (+0.5)
XSTNet-Base	audio-only*	SSL*	25.5	36.0	16.9	26.1 (+2.4)
XSTNet-Expand	MT, audio-only*	SSL*, MT	**27.8**	**38.0**	**18.4**	**27.8 (+4.1)**

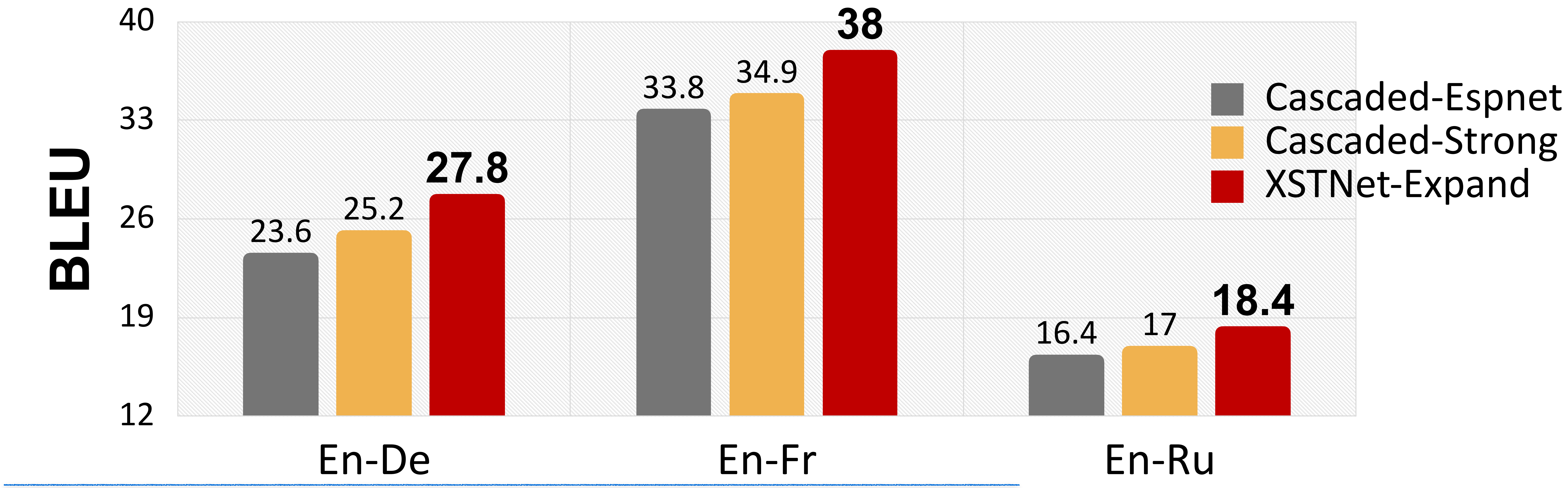
Table 2: *Performance (BLEU) on MuST-C En-De, En-Fr and En-Ru test sets. †: "Mono-data" means audio-only data from Librispeech, Libri-Light, as well as text-only data from Europarl/Wiki Text; *: "Audio-only" data from Librispeech audio data is used in the pre-training of wav2vec2.0-base module, and "SSL" means the self-supervised learning from unlabeled audio data.*

XSTNet-Base: Achieves the SOTA in the restricted setup

XSTNet-Expand: Goes better by using extra MT data

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

XSTNet better than cascaded ST! a gain of 2.6 BLEU



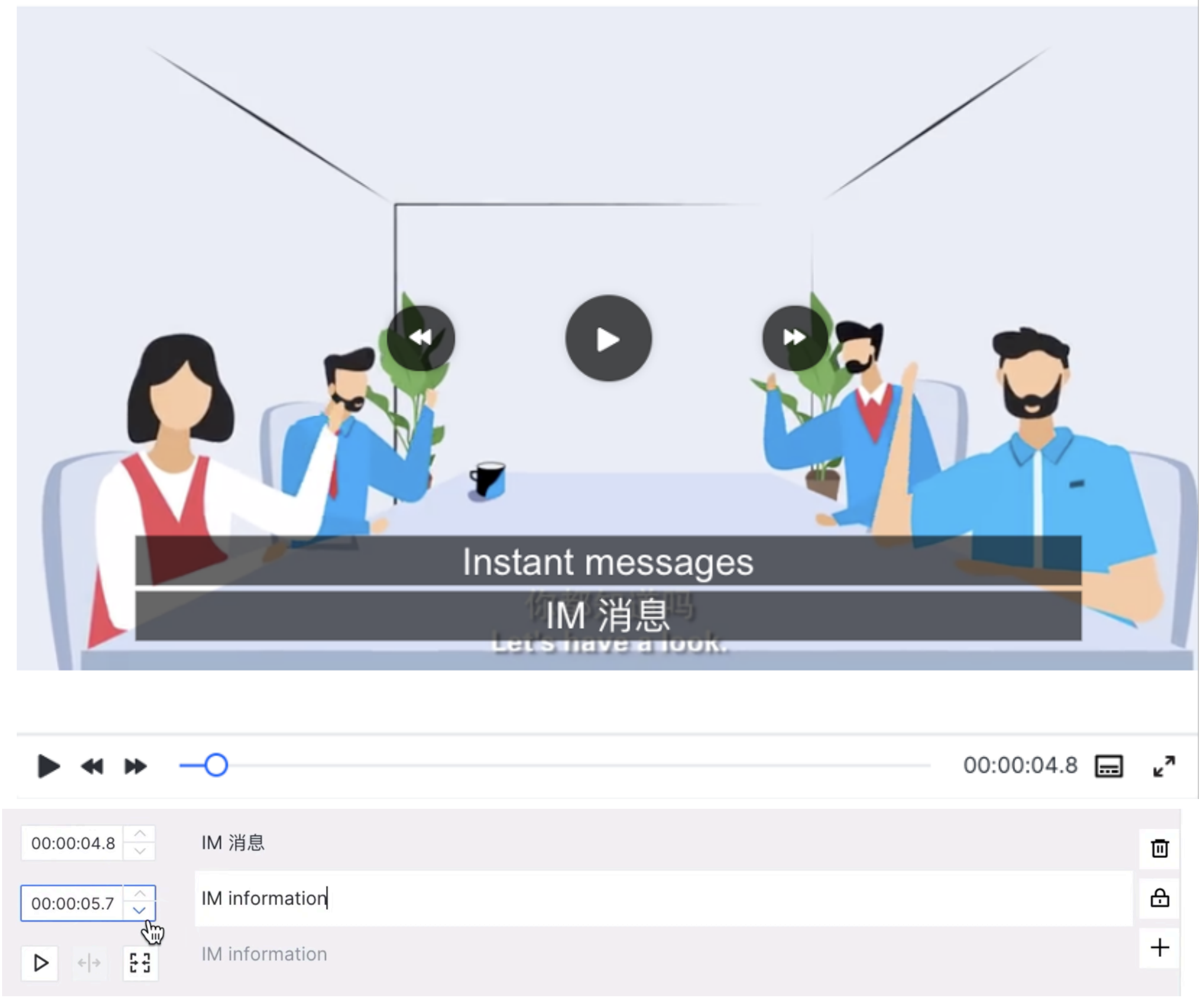
What is "Cascaded-Strong" system?

Strong ASR model + Large-scale MT data

Cascaded - Strong	Model	Training data	Performance (En-De)
ASR	W2V2+ Transformer	MuST-C D_{ASR}	WER=13.0
MT	Transformer-base	WMT + MuST-C D_{MT}	BLEU=31.7

VolcTransStudio: Video Translation Platform

火山翻译



Instant messages

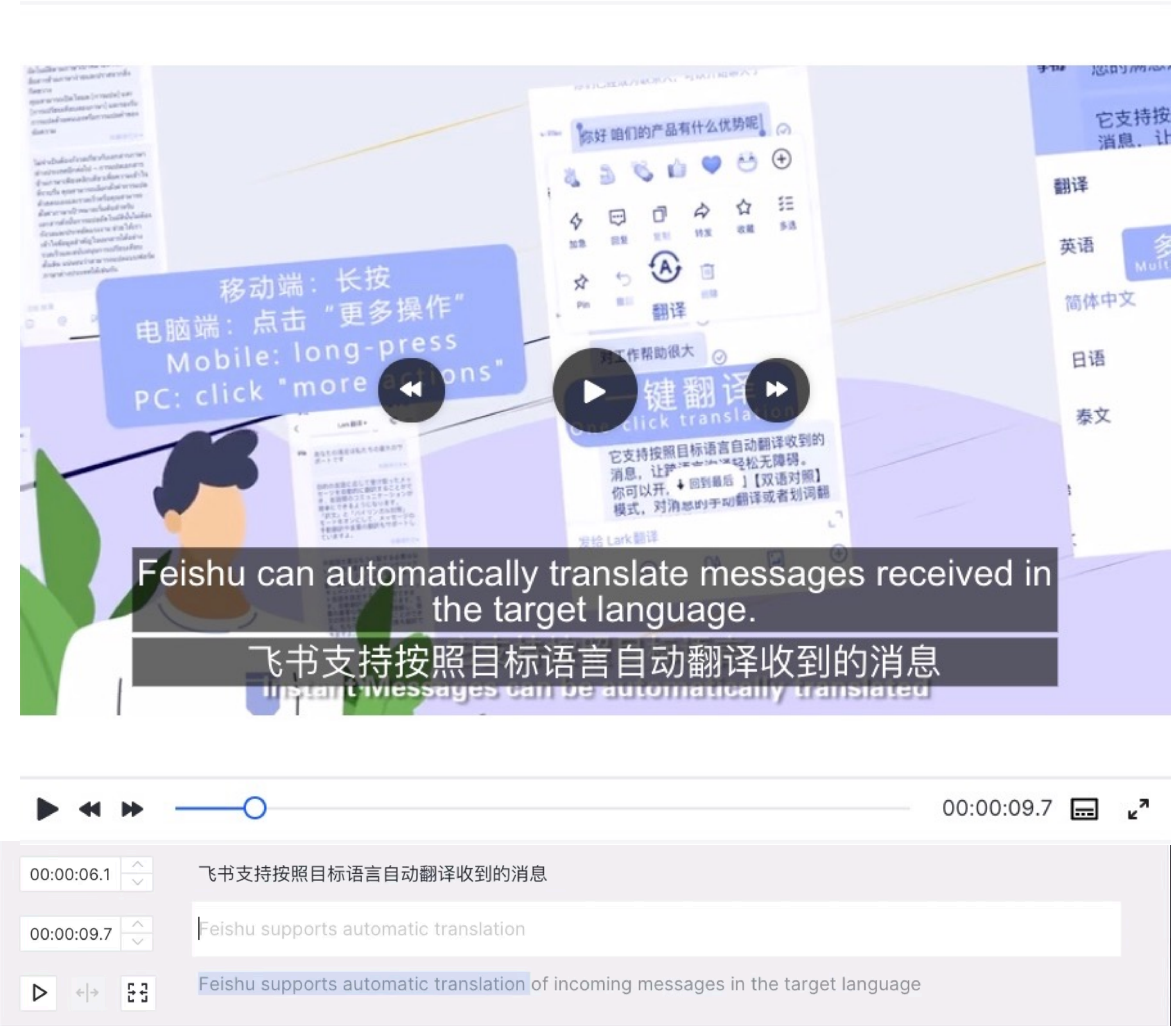
IM 消息

00:00:04.8

00:00:04.8 IM 消息

00:00:05.7 IM information

IM information



Feishu can automatically translate messages received in the target language.

飞书支持按照目标语言自动翻译收到的消息

00:00:09.7

00:00:06.1 飞书支持按照目标语言自动翻译收到的消息

00:00:09.7 Feishu supports automatic translation

Feishu supports automatic translation of incoming messages in the target language

实时翻译，自动提示 & 交互式修改

Correct-and-Memorize: Learning to translation from interactive revisions [Rongxiang Weng, Hao Zhou, Shujian Huang, Yifan Xia, Lei Li, Jiajun Chen. IJCAI 19]

Summary

- End-to-end Speech-to-Text works!
- Use external ASR, MT data, and audio/text for auxiliary signals
- Model
  - LUT: two-stage encoder, additional BERT KD [Dong et al AAAI 2021a]
  - Chimera: Shared semantic space encoder with fixed-size memory [Han et al ACL 2021]
  - COSTT: consecutive transcription-translation decoder [Dong et al AAAI 2021b]
- Training technique
  - Audio pre-training: Wave2Vec2.0[Baevski et al 2021]
  - External MT Pre-training
  - XSTNet: Progressive multi-task training [Ye et al Interspeech 2021]

Thanks

火山翻译官网

火山翻译公众号

NeurST neural speech translation toolkit

https://github.com/bytedance/neurst

LightSeq High performance sequence inference

https://github.com/bytedance/lightseq