CS11-737 Multilingual NLP

Speech Translation

Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

HW2 Update

- ASR for Quechua will be optional
  - bonus 20pts if completed
  - additional data: http://festvox.org/cmu_wilderness/
  - http://lrec-conf.org/workshops/lrec2018/W14/pdf/4_W14.pdf

Mid-term Report

- Everything in proposal with adjustment,
  - project description
  - data
  - evaluation procedure/metric
  - a baseline model and baseline results
  - Clearly state individual team member contribution
  - Use ACL paper template in latex

Speech-to-Text Translation(ST)

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

Application Type

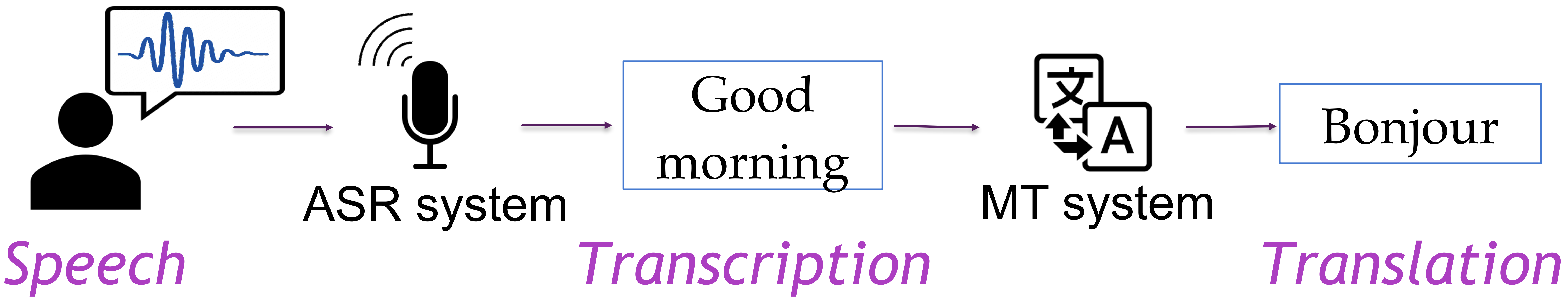
- (Non-streaming) ST e.g. video translation
- Streaming ST e.g. realtime conference translation

System

- Cascaded ST
- End-to-end ST

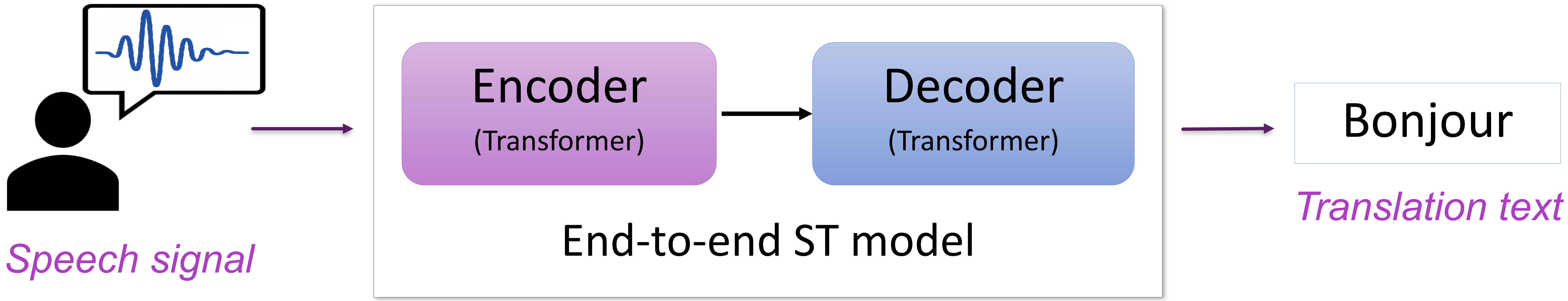
Cascaded ST System

- Challenges:
  1. **Computationally inefficient**
  2. **Error propagation**: Wrong transcription ➡️ Wrong 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



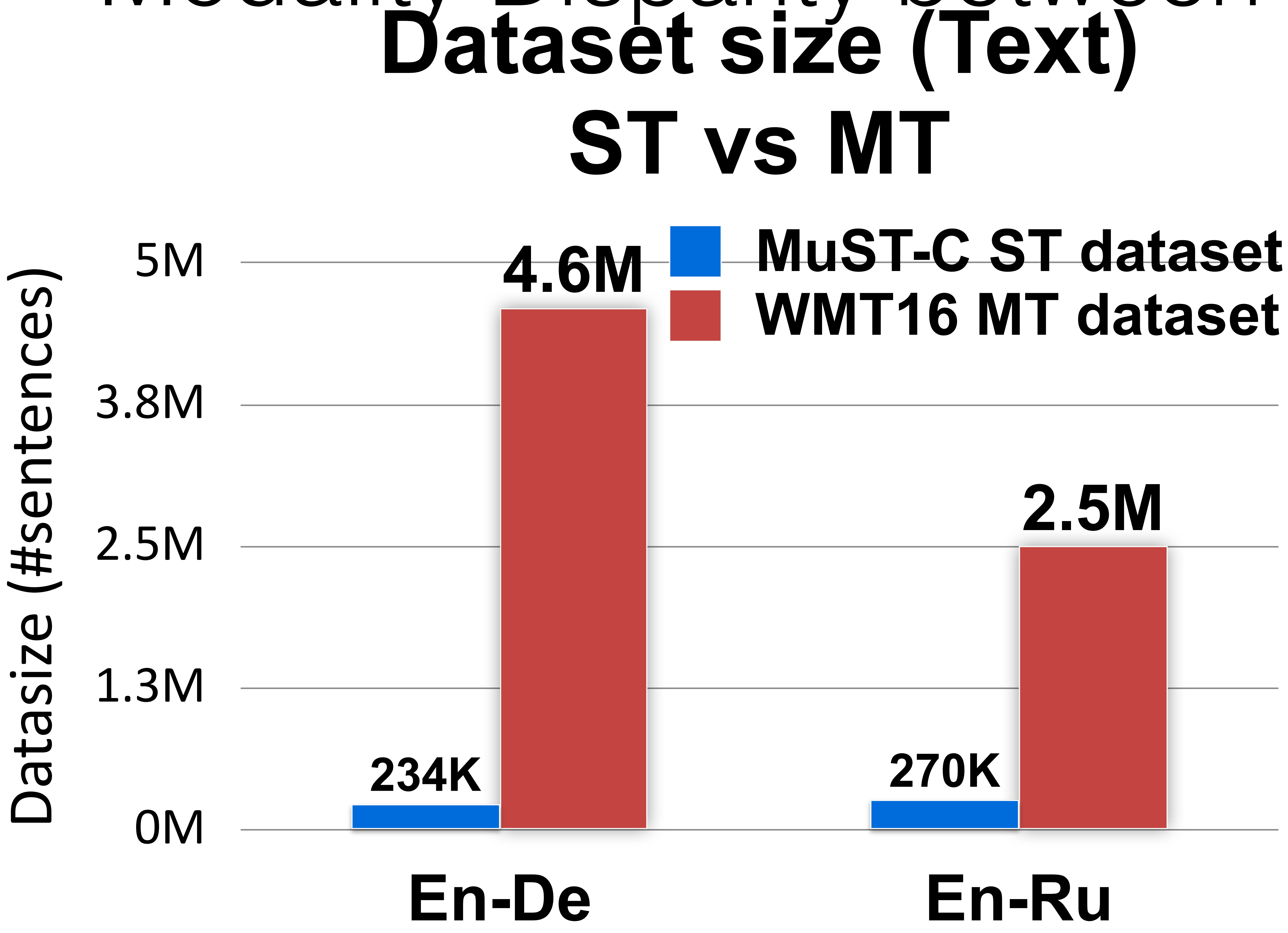
- Single model to produce text translation from speech
- Basic model: Encoder-Decoder architecture (e.g. Transformer)
- Advantage:
  - Reduced latency, simpler deployment
  - Avoid error propagation

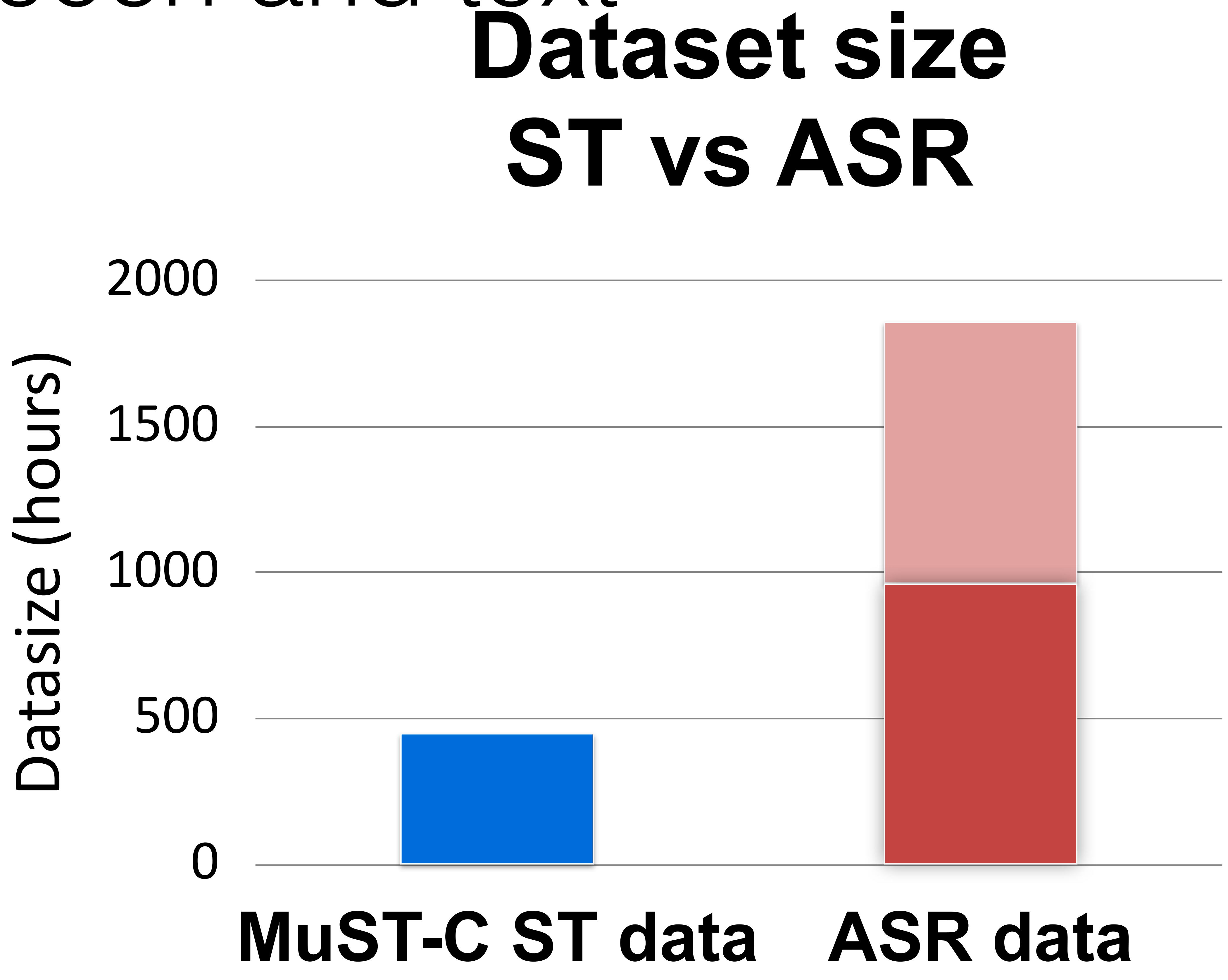
[1] Bérard et al., Listen and translate: A proof of concept for end-to-end speech-to-text translation. 2016

Challenge

- Data scarcity - lack of large parallel audio-translation corpus

- Modality Disparity between speech and text



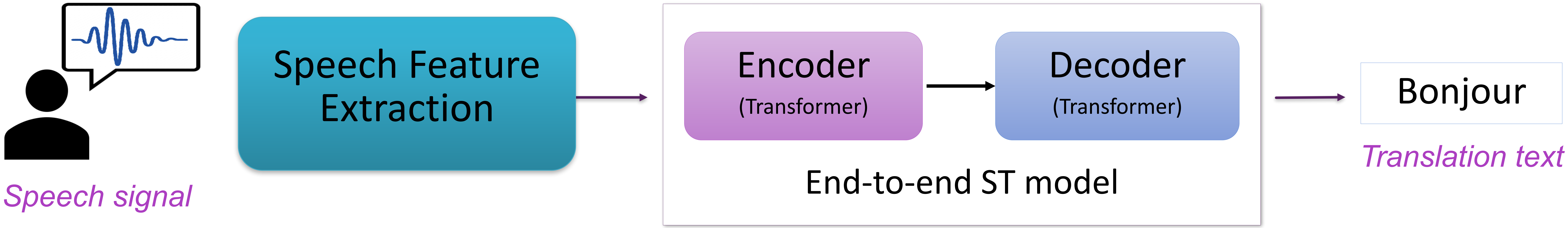


Challenge

- Modality Disparity between speech and text
  - Disfluencies
    - Hesitations: “uh”, “uhm”, “hmm”,
    - Discourse markers: “you know”, “I mean”,…
    - Repetitions: “It had, it had been a good day”
    - Corrections: “no, it cannot, I cannot go there”
  - Unlike (Text) MT, No punctuation
    - let s eat grandpa
    - Let’s eat, Grandpa !

Basic End-to-end ST Model

Basic ST model

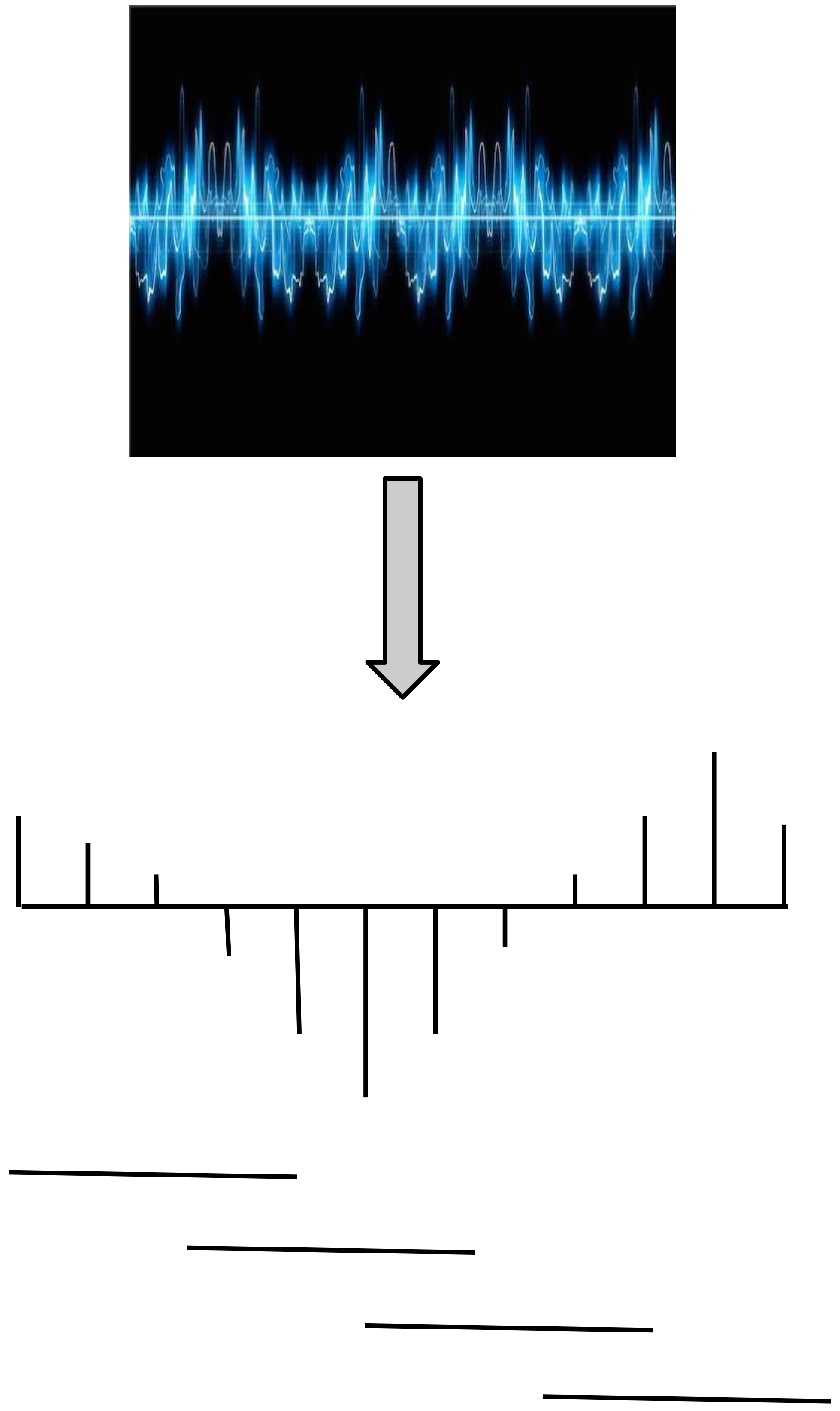


Main differences to text machine translation

Input: Audio signal are continuous and much longer!

Audio Signal - Same as ASR

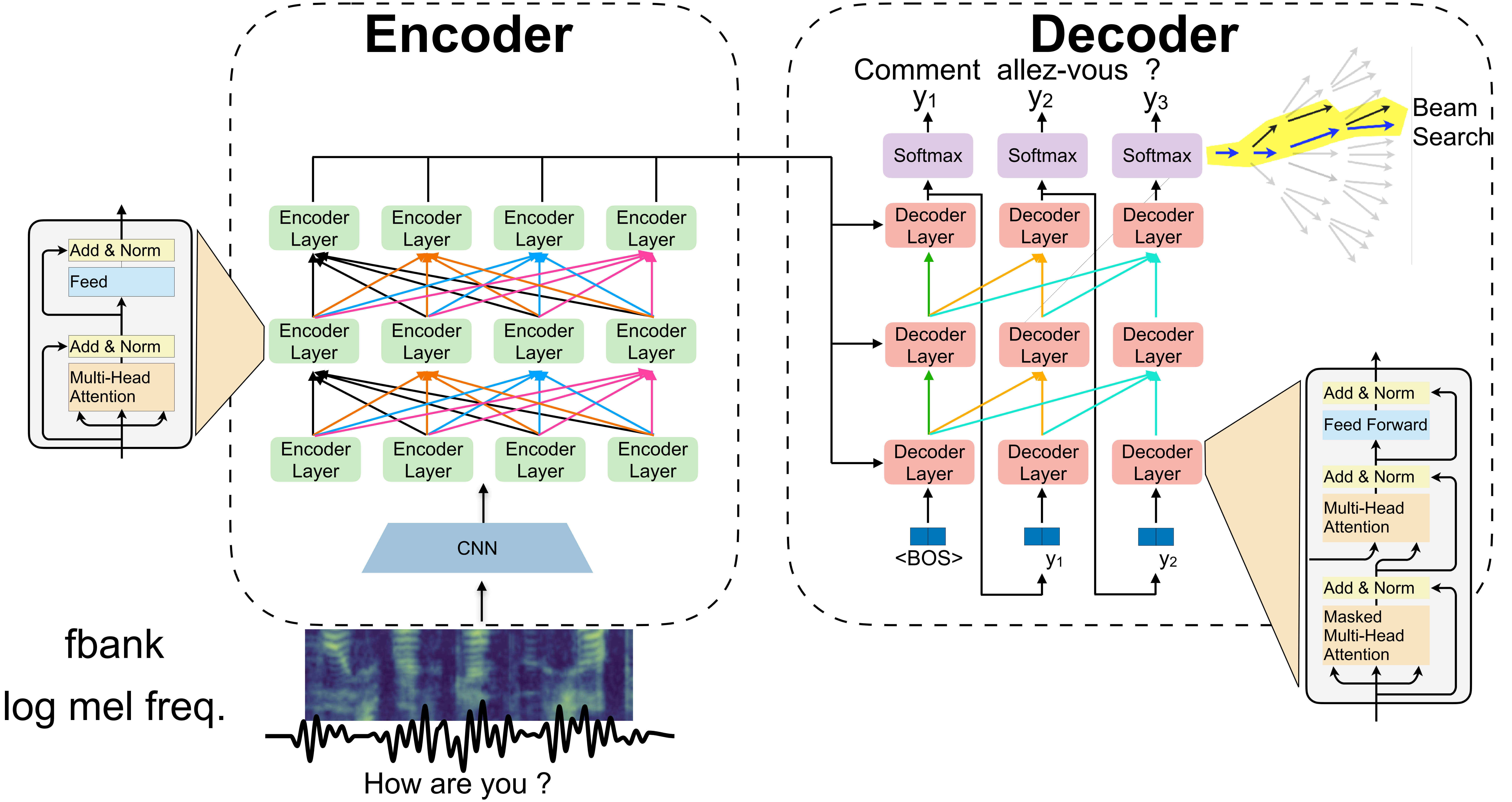
- Following best-practice from ASR
- Signal Sampling
  - Measure Amplitude of signal at time t
  - Typically 8kHz or 16 kHz
- Windowing — Frame
  - Split signal in different windows, called Frame
    - Length: ~ 20-30 ms (typically 25ms)
    - Stride: ~ 10 ms



Basic Speech Translation Model (Similar to MT)

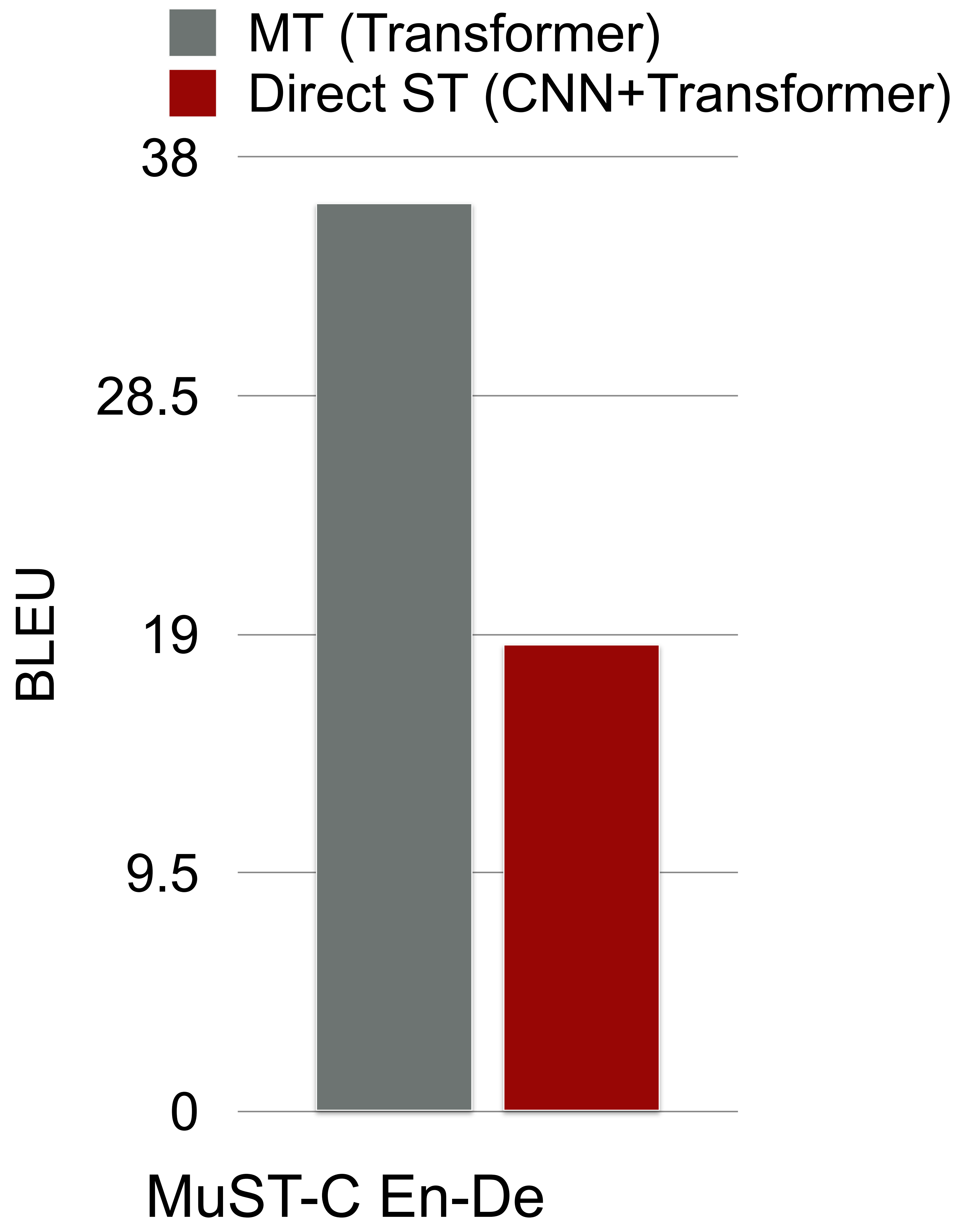
Transformer-based: N-layer convolution + attention encoder, M-layer decoder

Training data: <audio seq., translation text>



Speech Translation model lags behind MT

- Performance on MuST-C En-De:
  - ST 18.6
  - MT 36.2 (taking correct transcript as input)

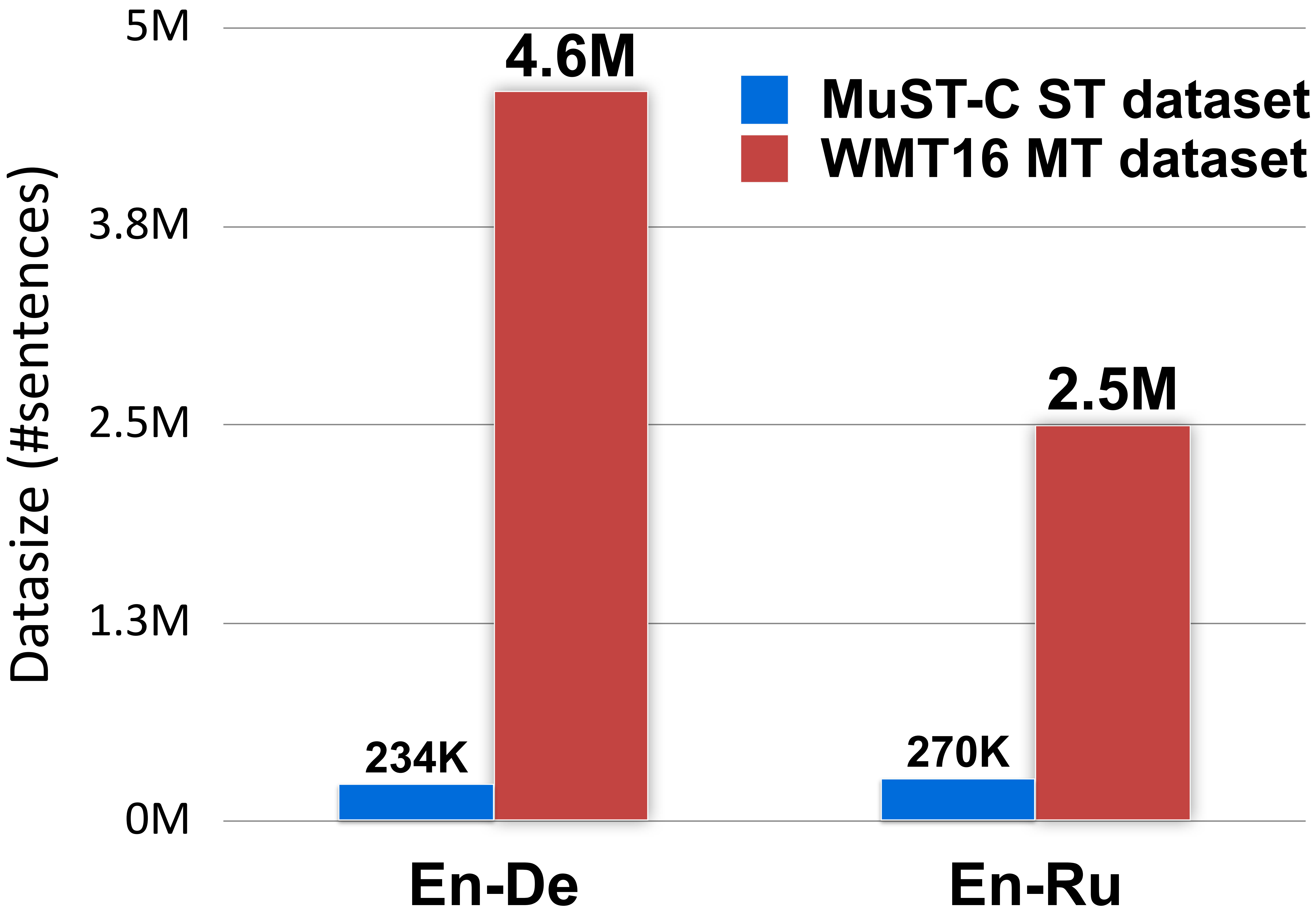


Approaches for Speech Translation

- Utilizing additional parallel text from MT corpus - MT pretraining
  - Decoder initialization from separately trained MT model
  - Single-modal(audio) Encoder-Decoder: COSTT[Dong et al, AAAI 2021b]
- Using Additional ASR data - ASR Pre-training
  - Curriculum Pre-training [Wang et al, ACL 2020]
  - LUT [Dong et al, AAAI 2021a]
- Using additional raw audio data
  - Wav2vec & Wav2Vec2.0 [Schneider et al. Interspeech 2019, Baevski et al NeurIPS2020]
  - Apply to ST [Wang et al, 2021, Zhao et al, ACL 2021, Wang et al, Interspeech 2021]
- Distilling knowledge from Pre-trained Language Model (BERT)
  - LUT [Dong et al, AAAI 2021a]
- Learning Better Speech-text cross-modal representation for ST
  - TCEN-LSTM [Wang et al, AAAI 2020]
  - Chimera [Han et al, ACL 2021a]
  - Wav2vec2.0 + mBart + Self-training [Li et al, ACL 2021b]
  - FAT-ST [Zheng et al, ICML 2021]
  - ConST [Ye et al, 2022]
  - WACO [Ouyang et al, 2023]
- Better Fine-tuning Strategy
  - XSTNet [Ye et al, Interspeech 2021]

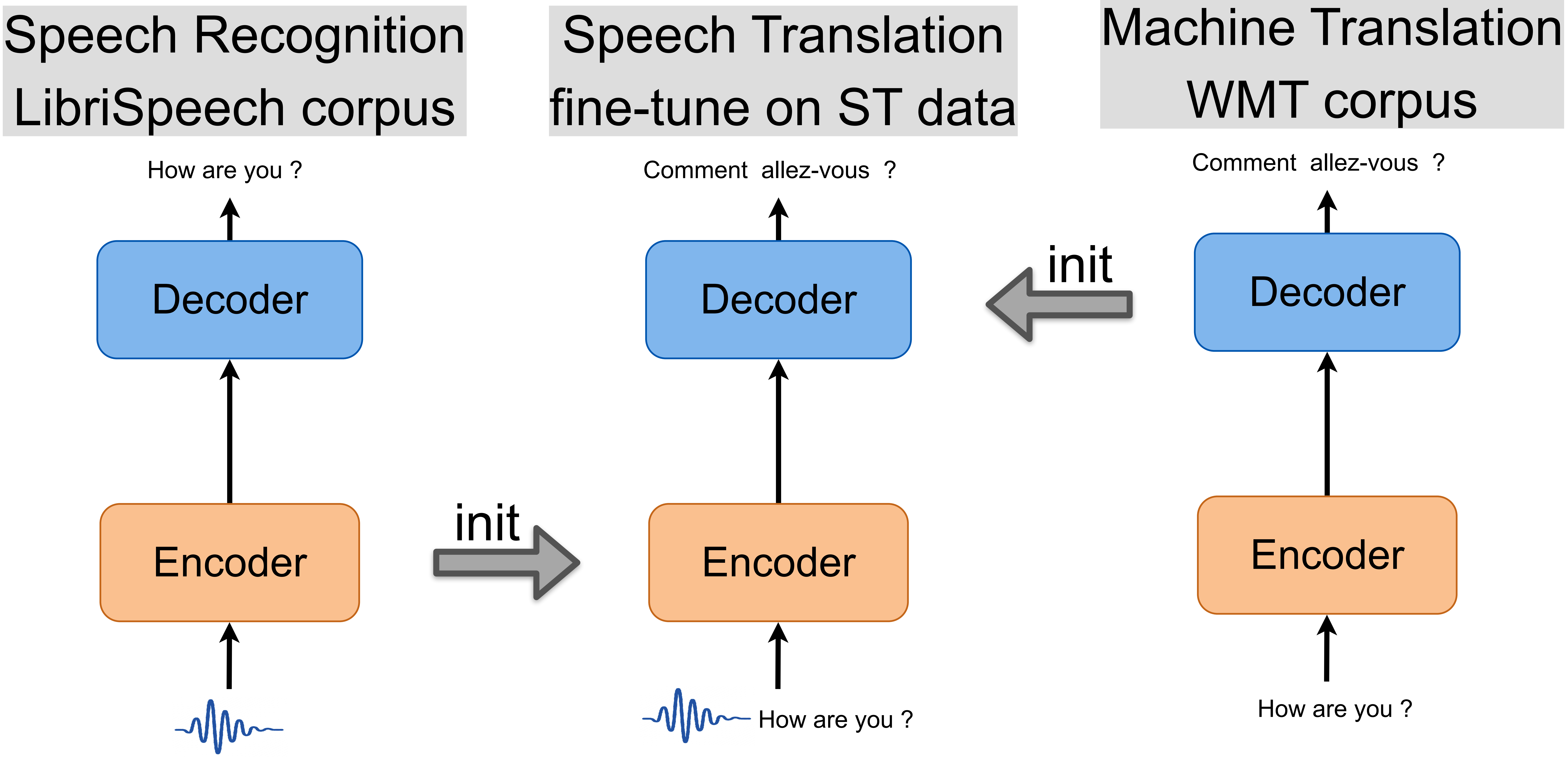
Using external Parallel Text

Dataset size
ST vs MT

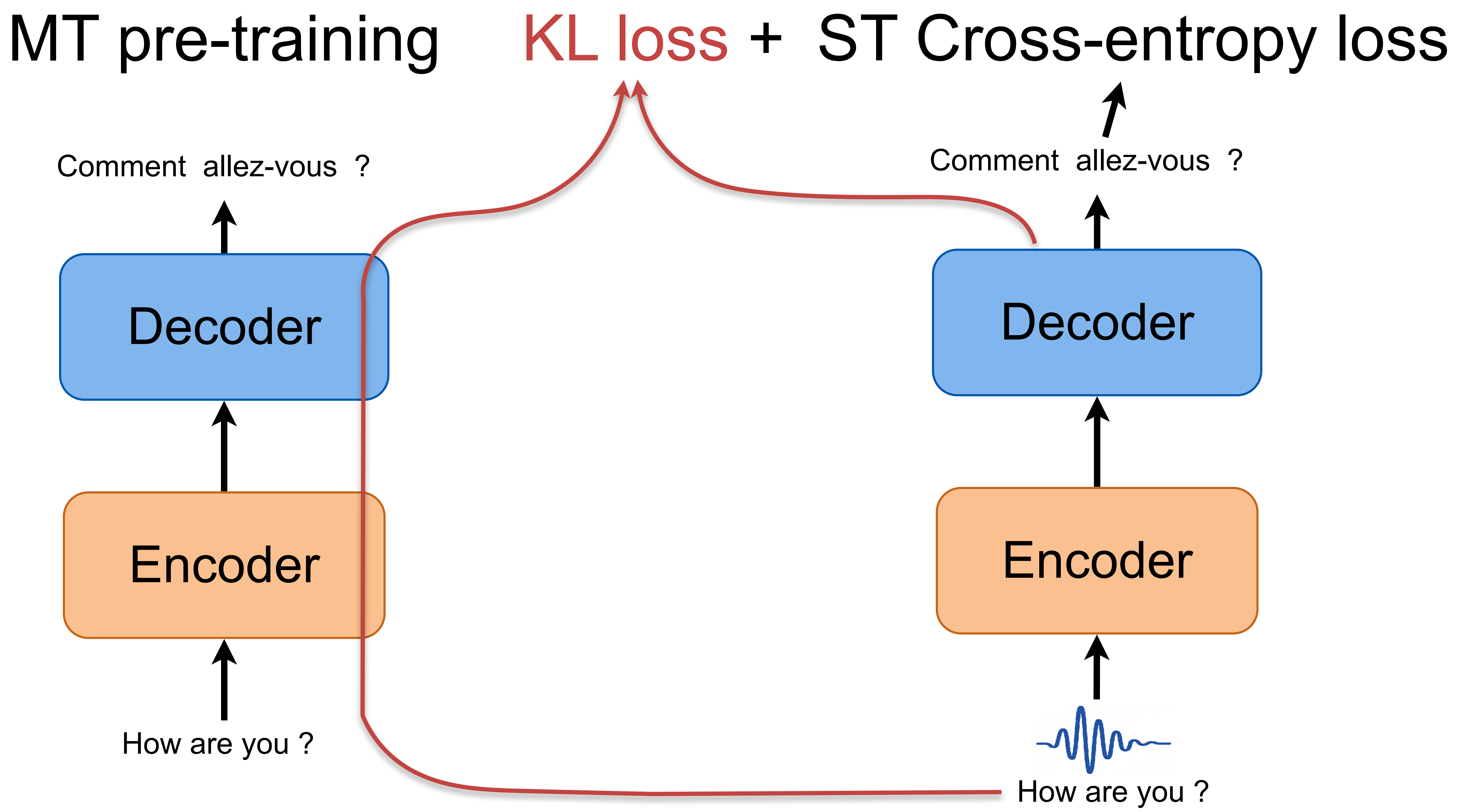


How to use MT data *with much larger scale* to improve ST performance?

Separate Encoder-Decoder Pre-train



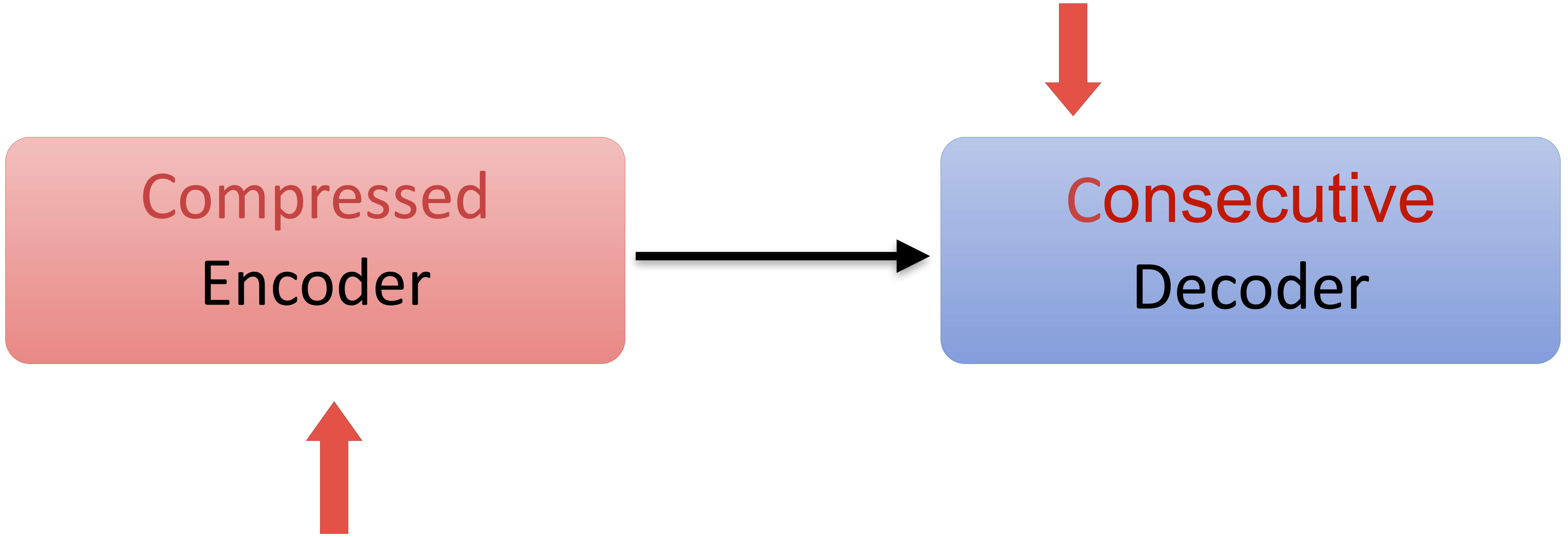
Knowledge Distillation from MT model



End-to-End Speech Translation with Knowledge Distillation [Liu et al, Interspeech 2019]

Motivation of Better Decoding

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



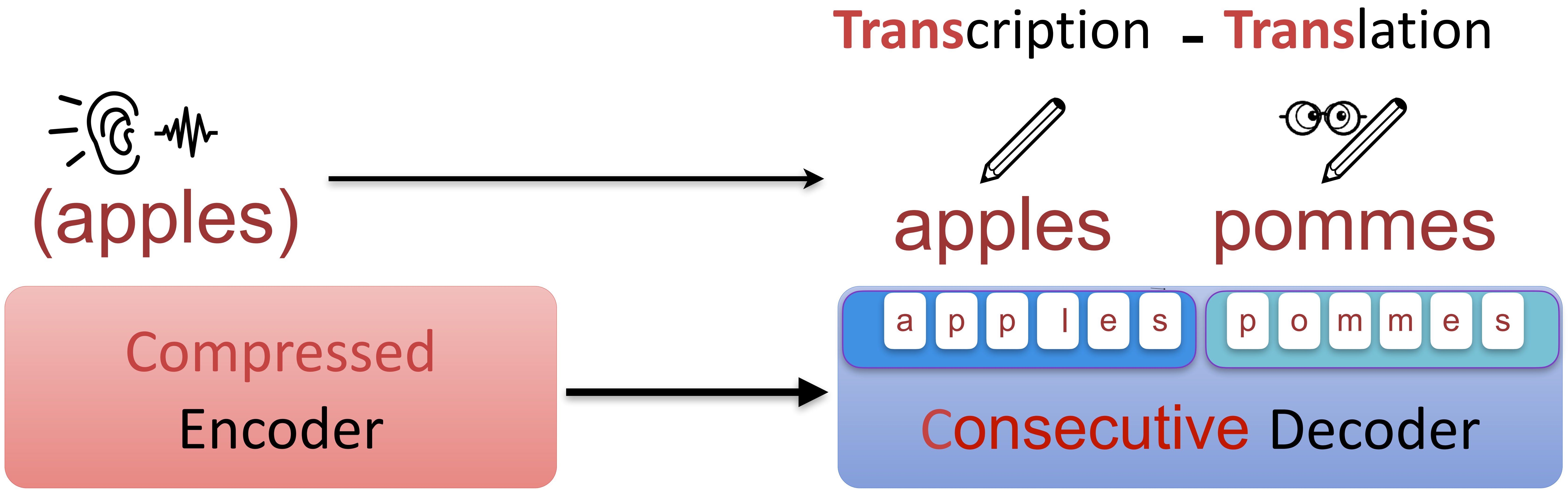
Problem2: Long acoustic sequence is challenging for the encoder!
Idea 2: Introduce a compressed encoder to relief the model memory.

Pre-train ST's decoder with full MT

How to make a single model's decoder to perform text translation?

Decoder ==> translation

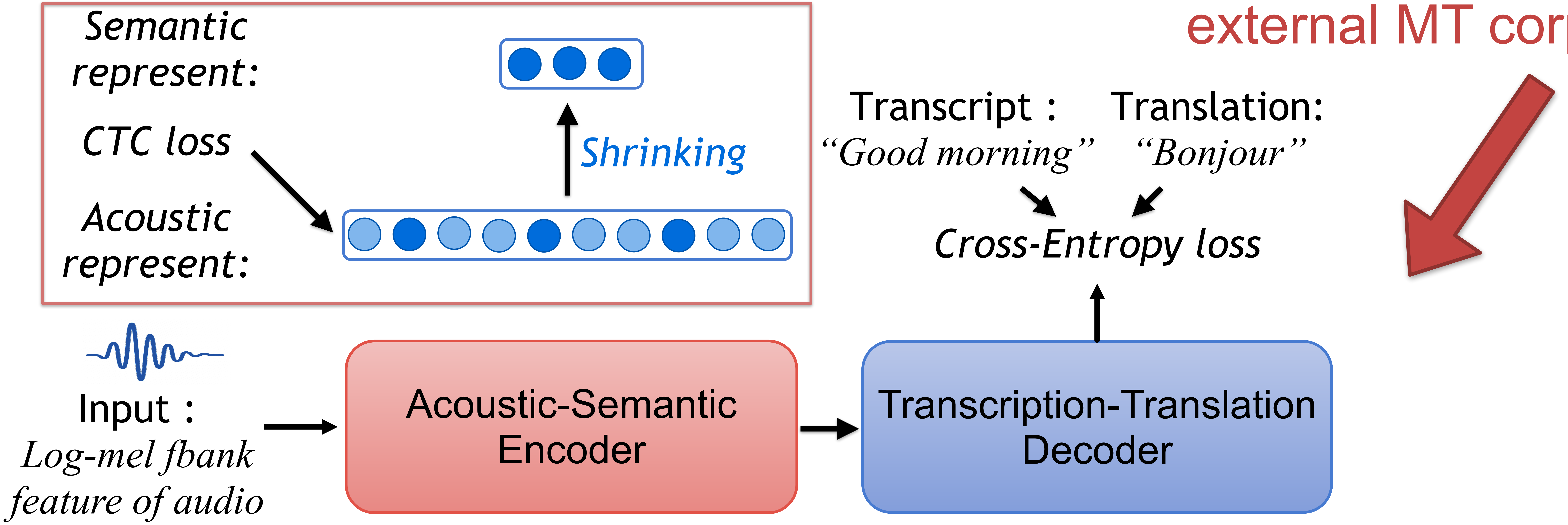
Encoder -> Decoder ==> transcribe and translation



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

COSTT for ST

Step1: Pre-train using external MT corpus



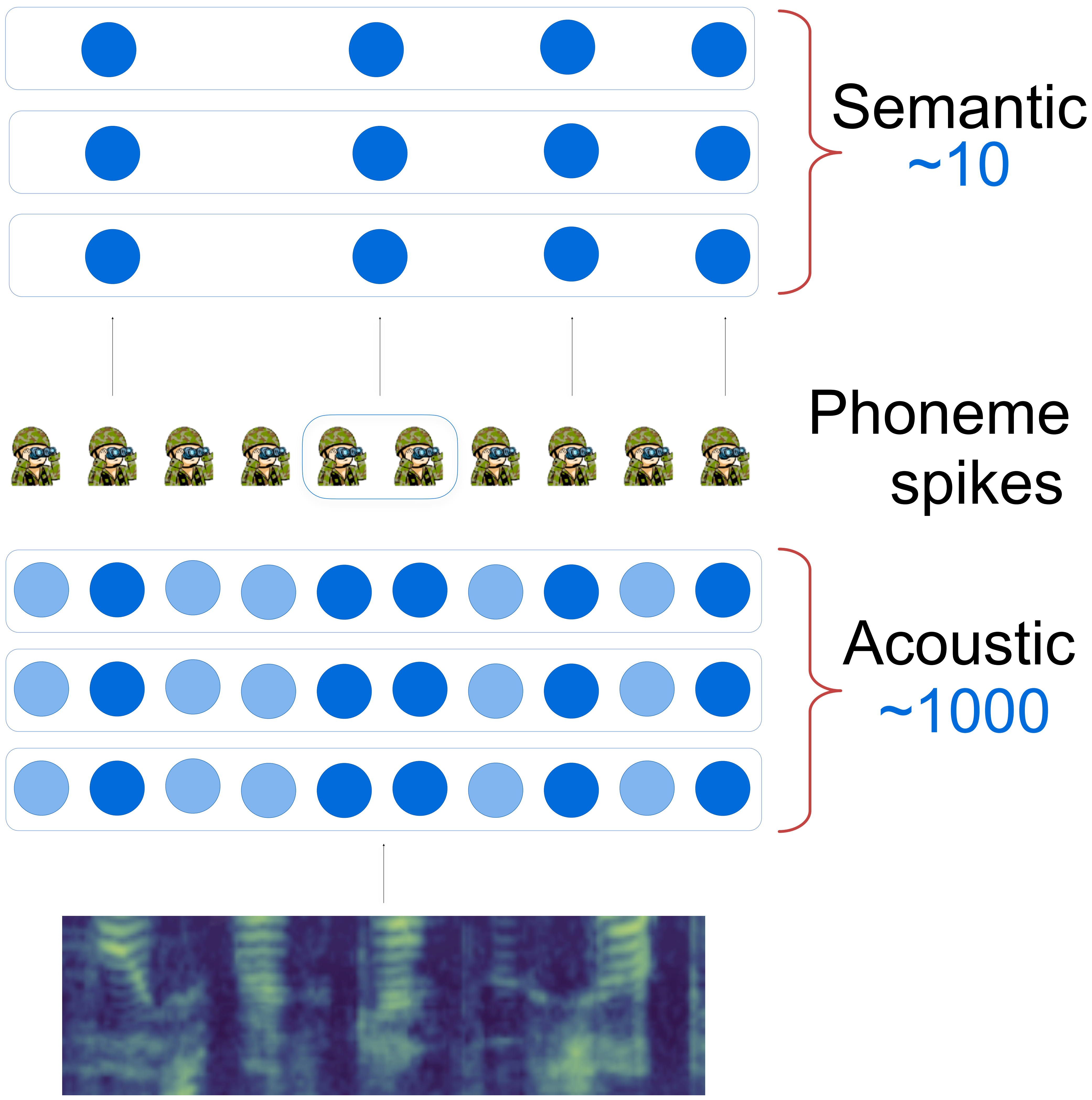
Step 2: Train encoder w/ shrinking module using CTC

Step 3: Train full model on ST data <audio, transcript, translation>

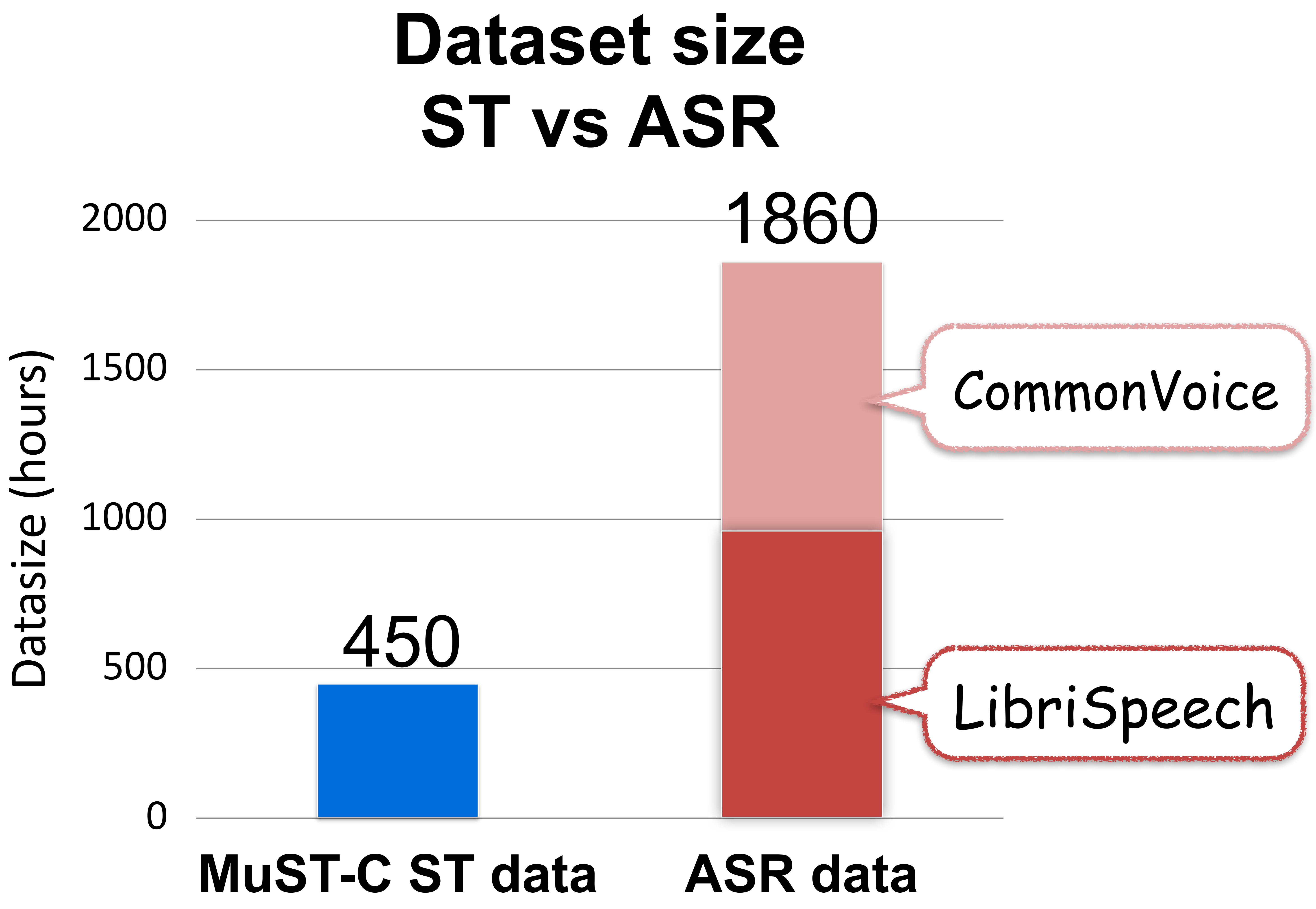
Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

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



Using external ASR data

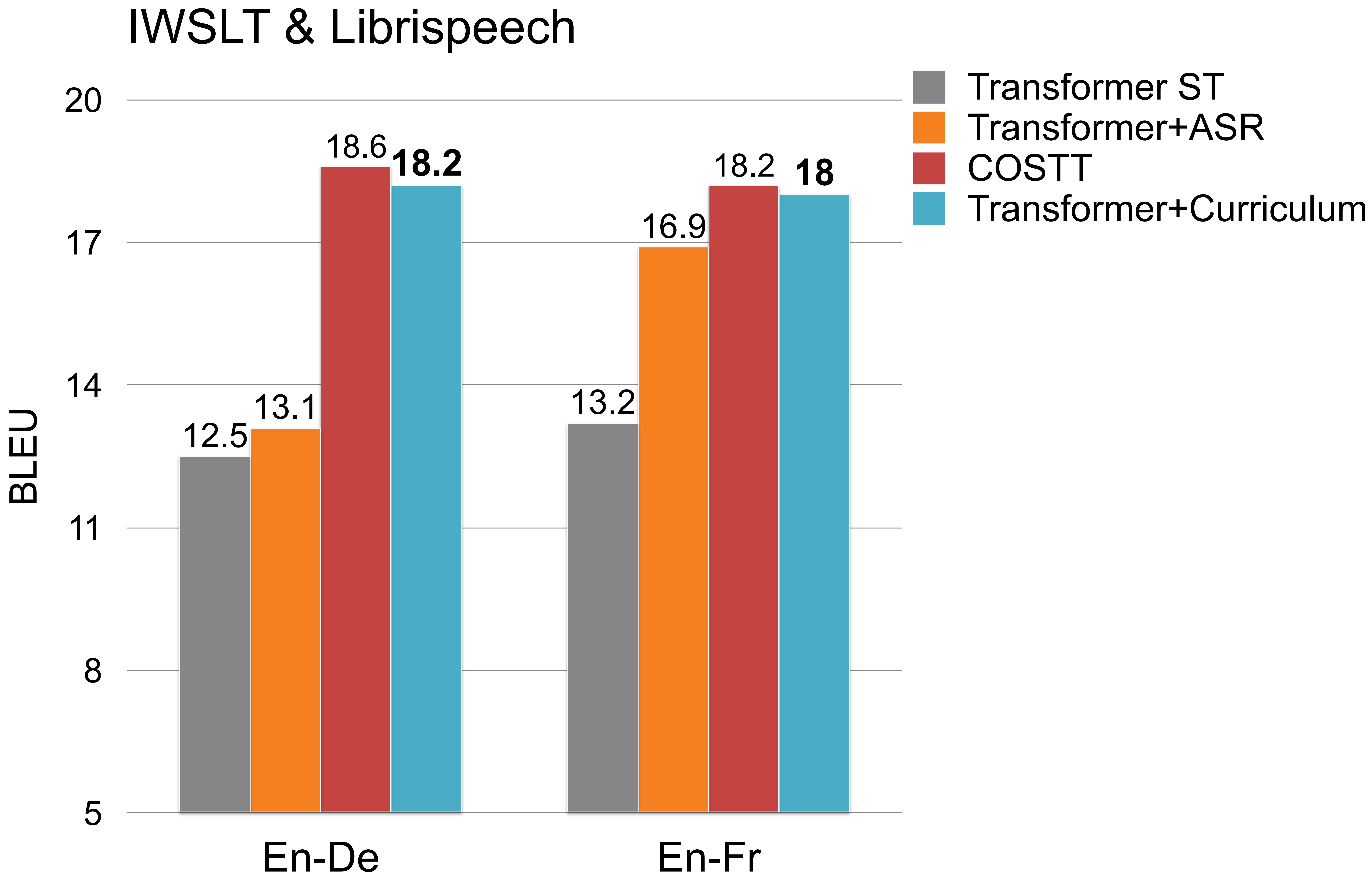


How to use larger external <u>ASR data</u> to improve ST performance?

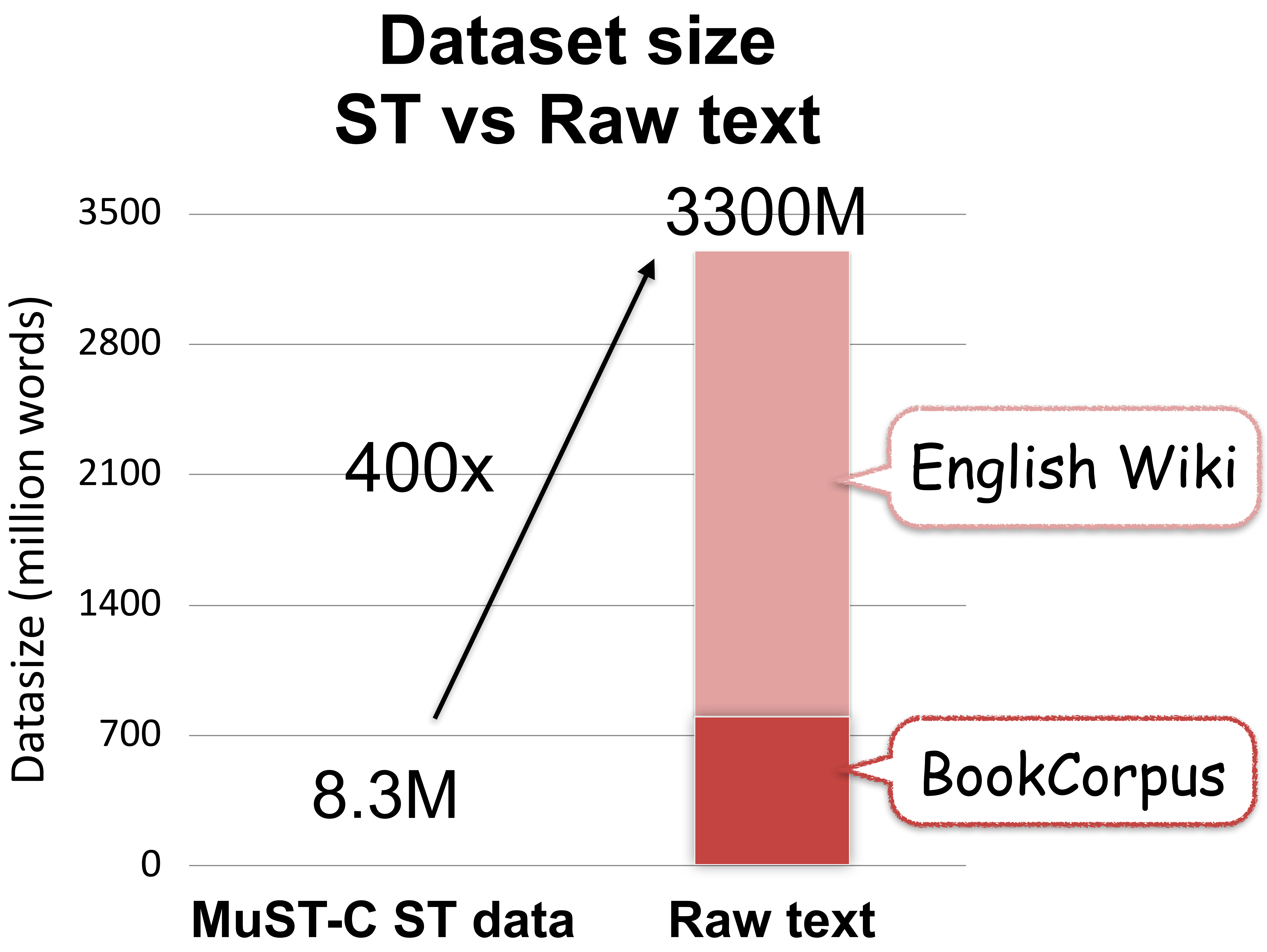
Curriculum Pre-training with ASR data

Curriculum Pre-training for End-to-end Speech Translation [Wang et al, ACL 2020]

ASR Pre-training helps ST



Raw Text Pre-training

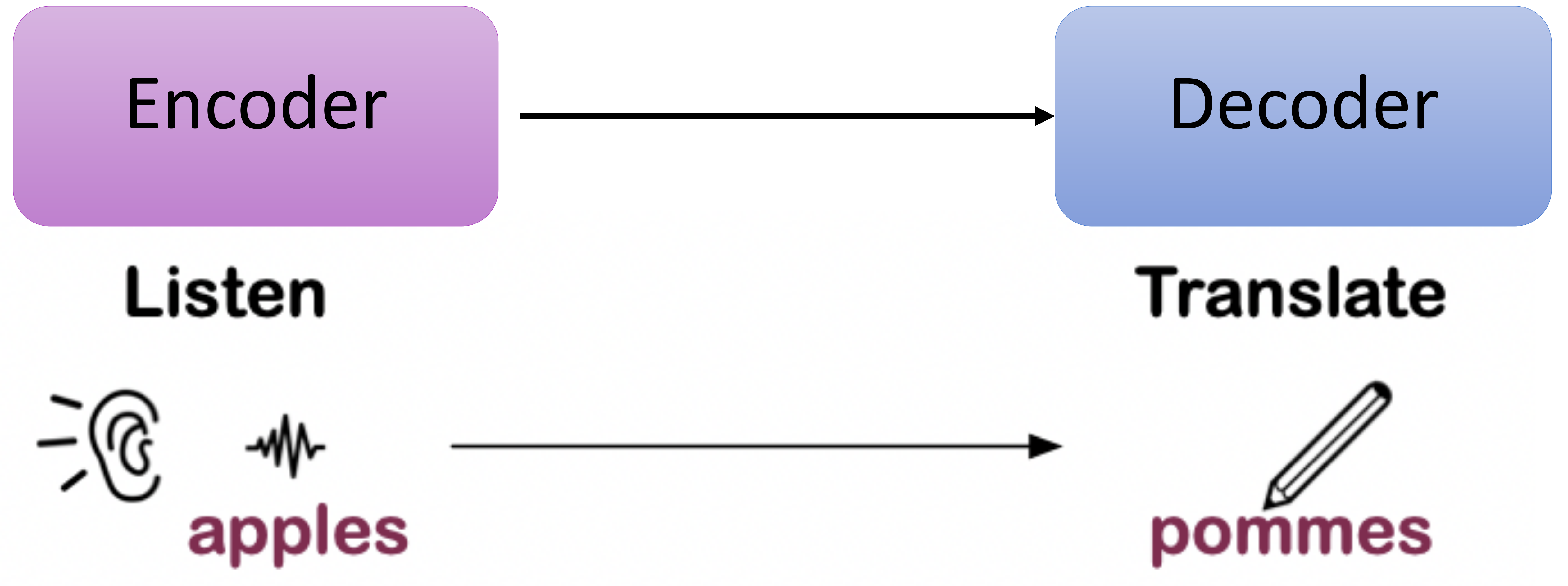


Using pre-trained LM in decoding weighting is easy!

But

🤔 How to use pre-trained BERT to improve ST performance?

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

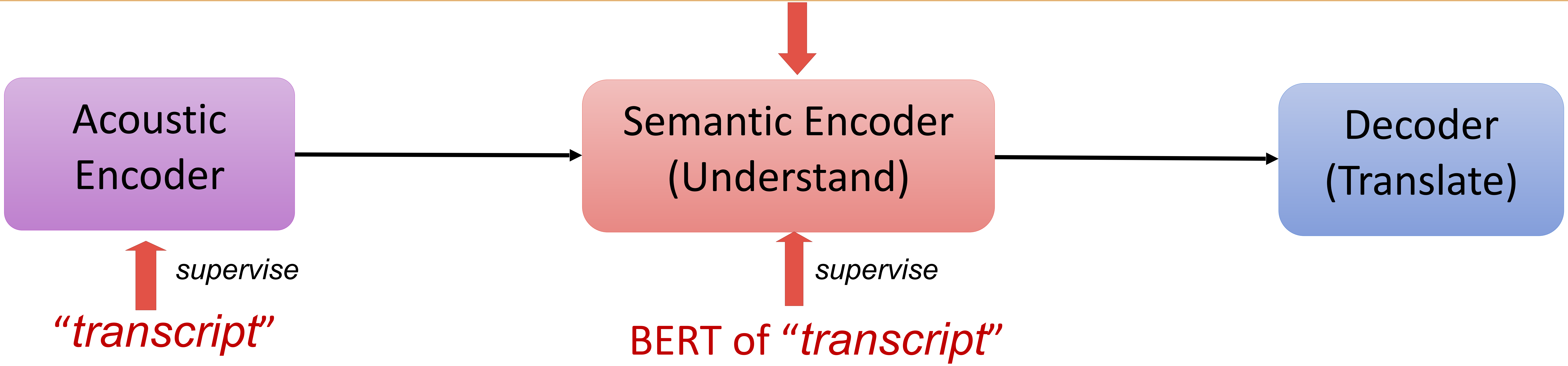
Question: How human translate?

"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



Drawback 2: Limit in using "transcript" info.

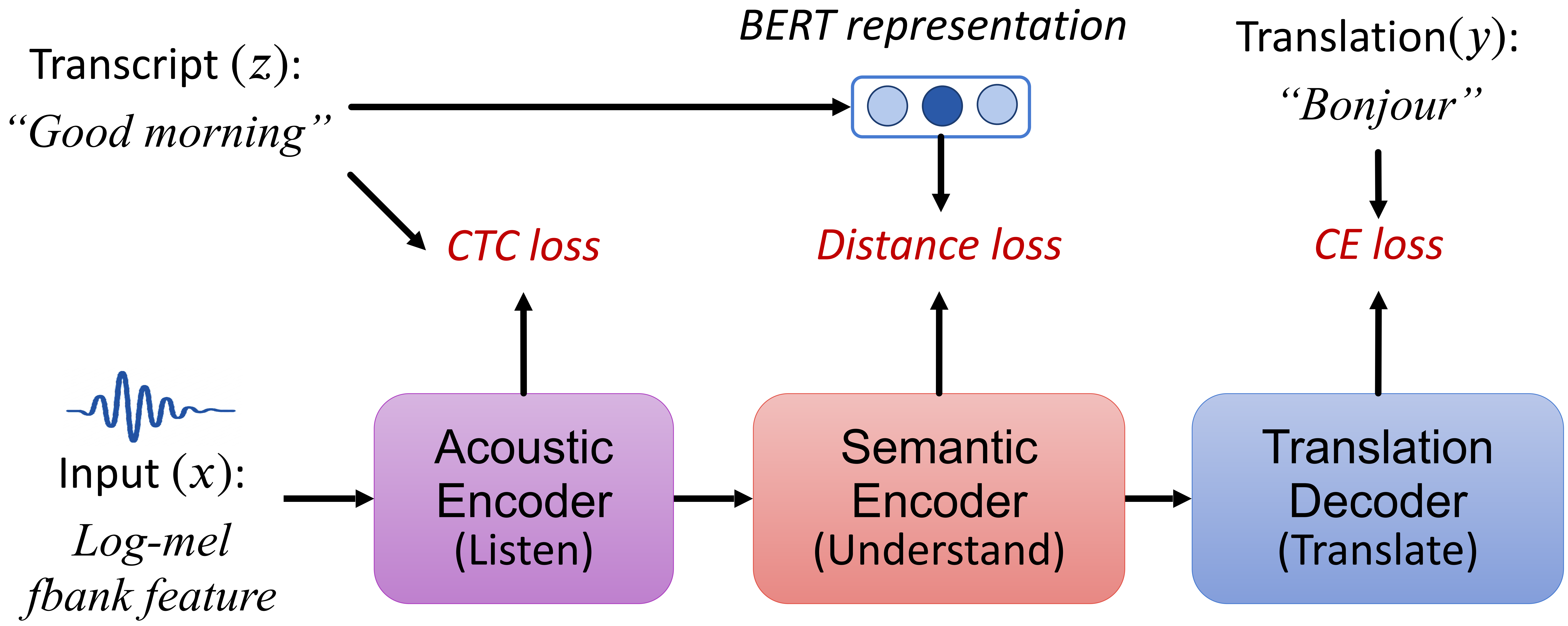
Idea 2: Utilizing the pre-trained representation (e.g. BERT) of the "transcript" to learn the semantic feature.

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

LUT: Utilizing Pre-trained Model on Raw Text

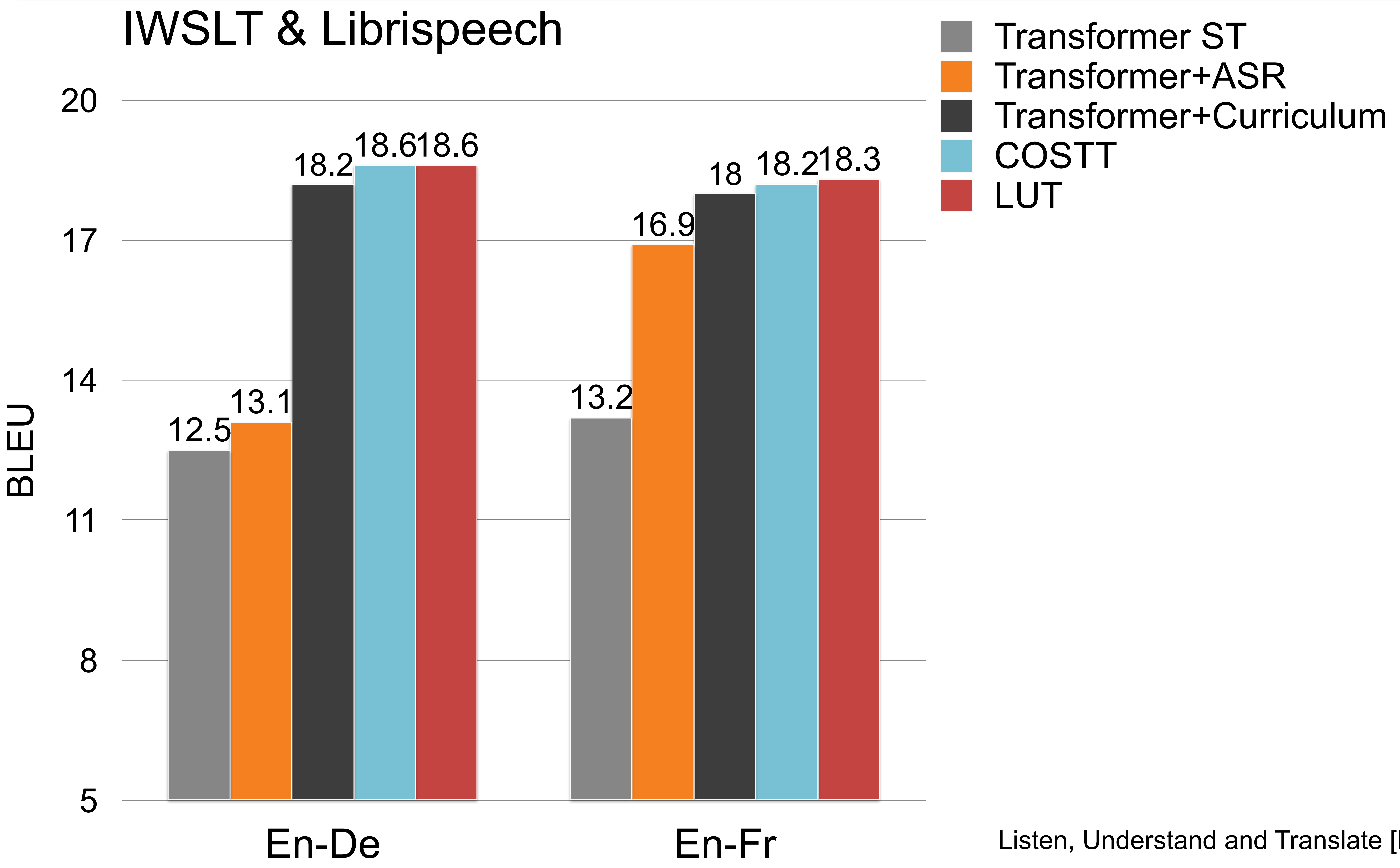
Training data: triples of

<speech, transcript_text, translate_text>



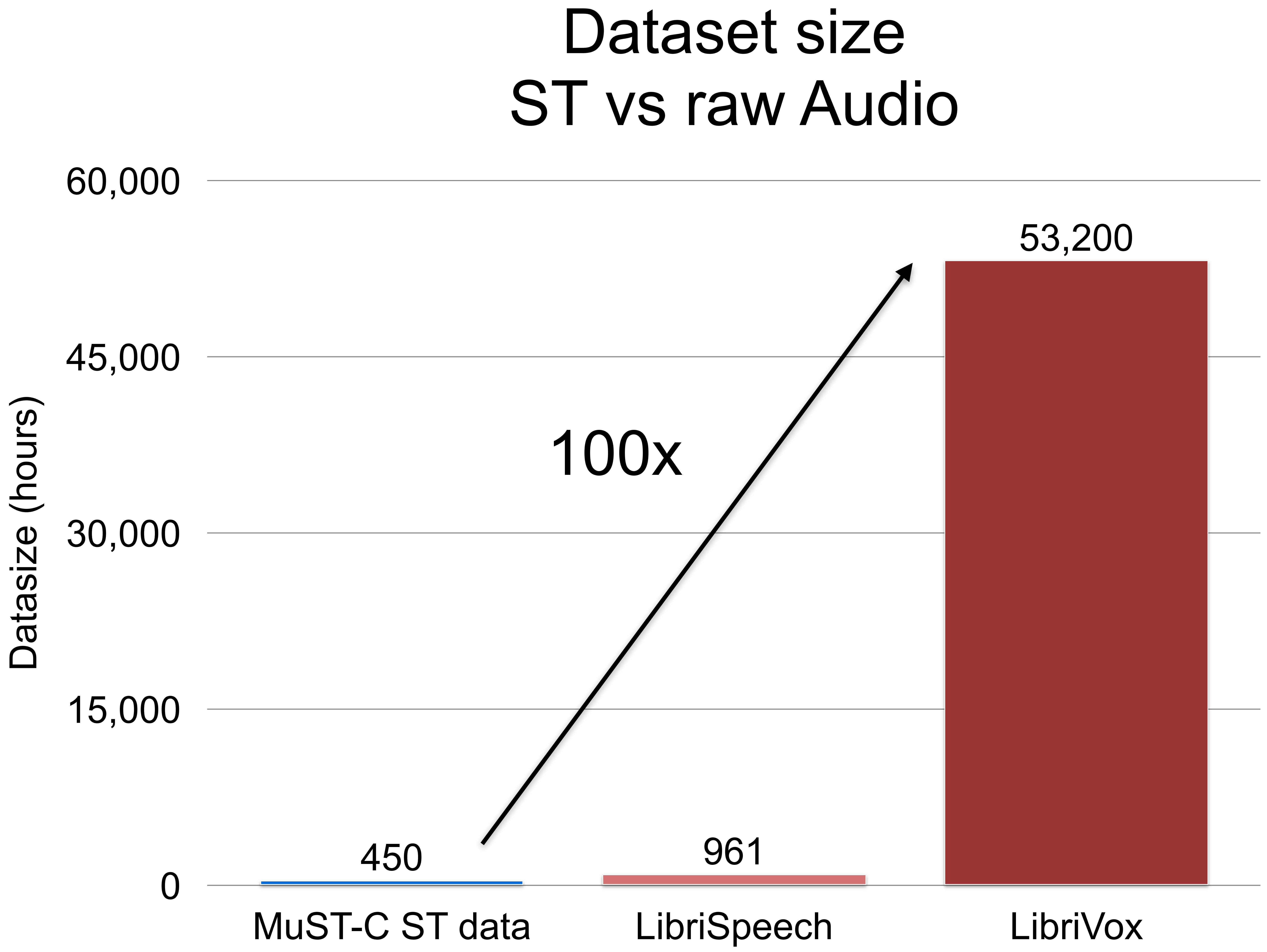
Listen, Understand and Translate [Q. Dong, R. Ye, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

ST Benefits from BERT, with Raw Text Pre-training



Listen, Understand and Translate [Dong et al, AAAI 2021]

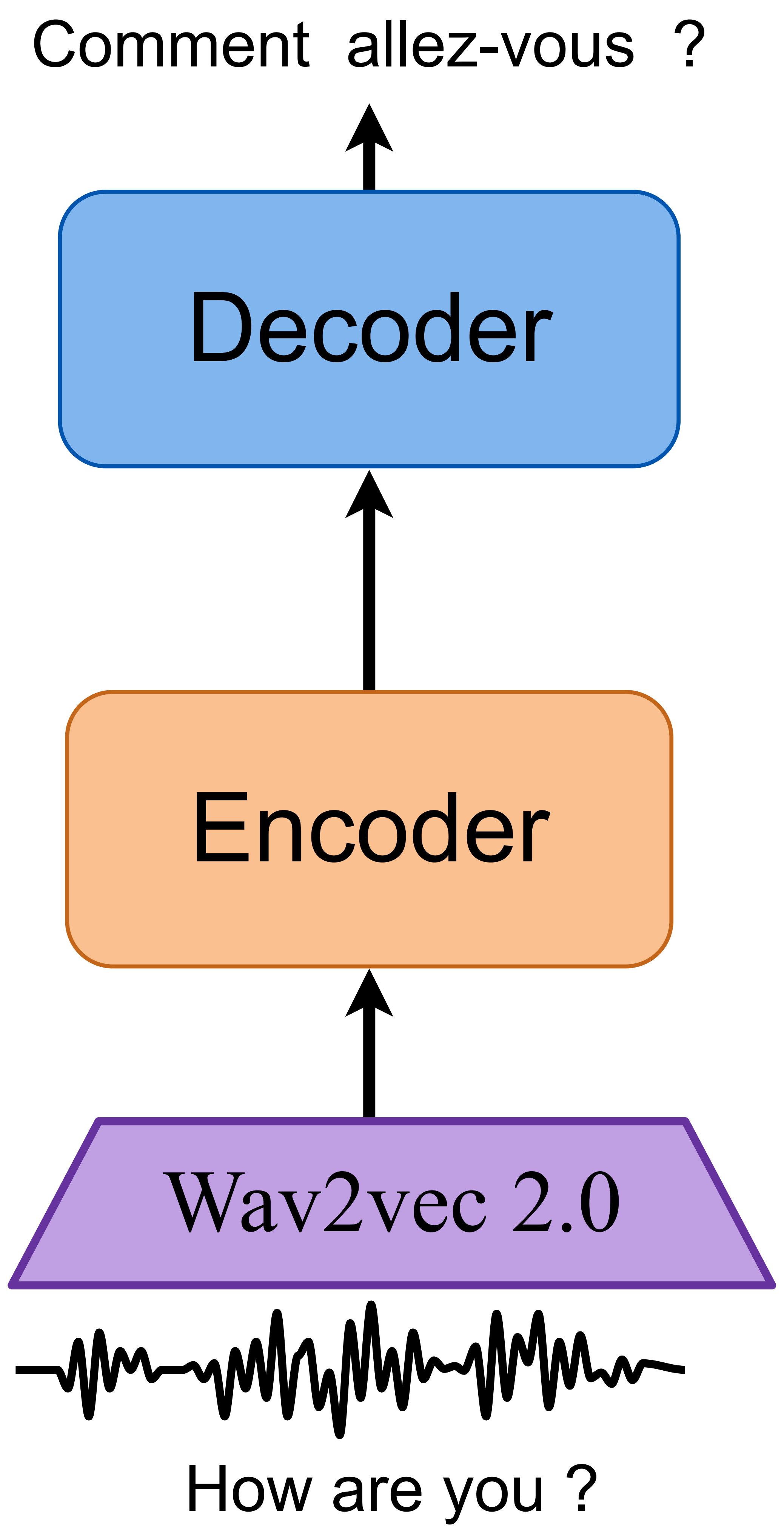
Audio Pre-training

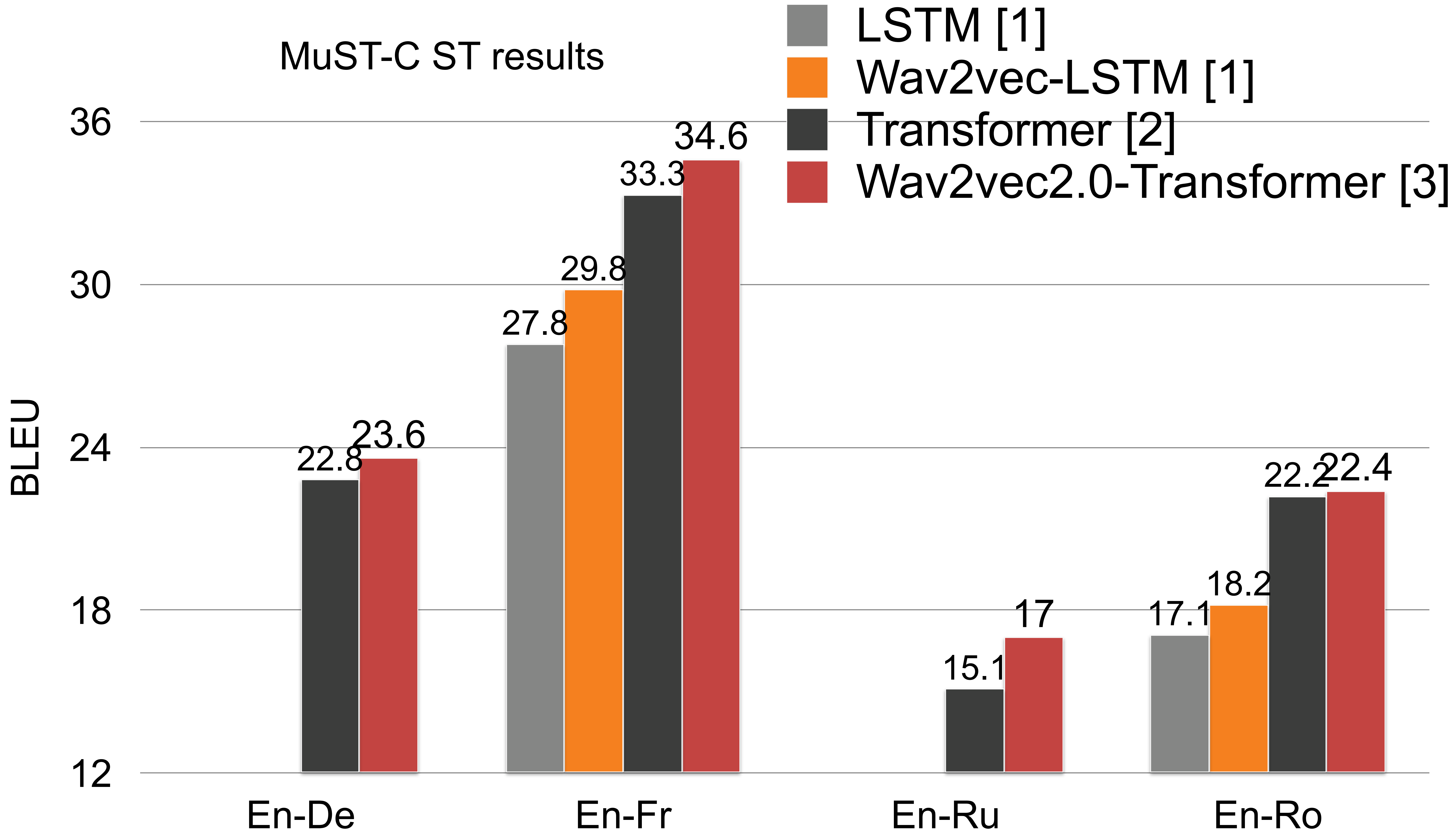


🤔 How to use larger raw audio data to improve ST performance?

Speech Translation with Audio-Pretrain

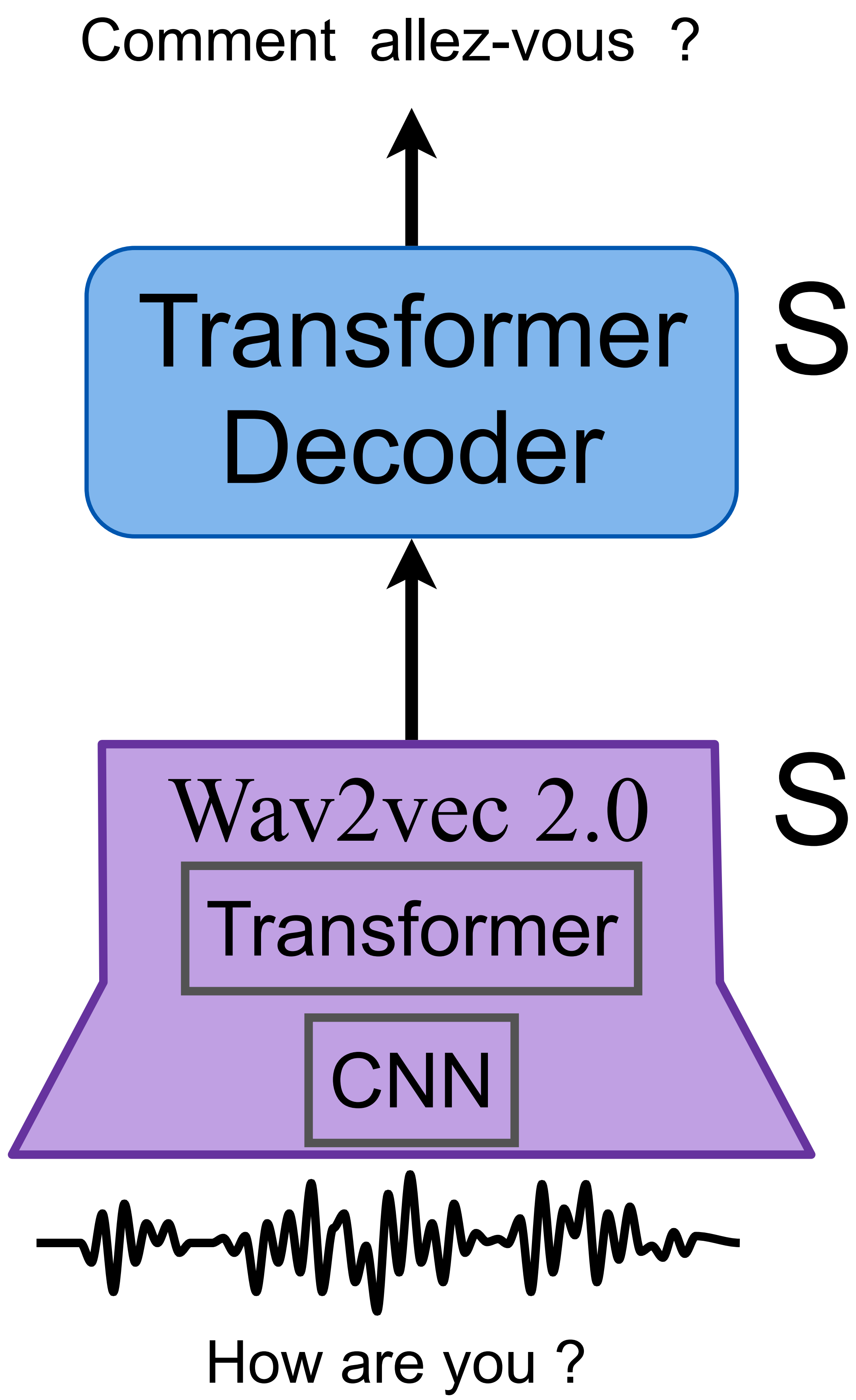
Wav2vec Pretrain + Fine-tune on ST





[1] Self-supervised Representations improve end-to-end speech translation [Wu et al. InterSpeech 2020]
[2] NeurST toolkit [Zhao et al ACL2021 demo] [3] End-to-end Speech Translation [Ye et al. InterSpeech 2021]

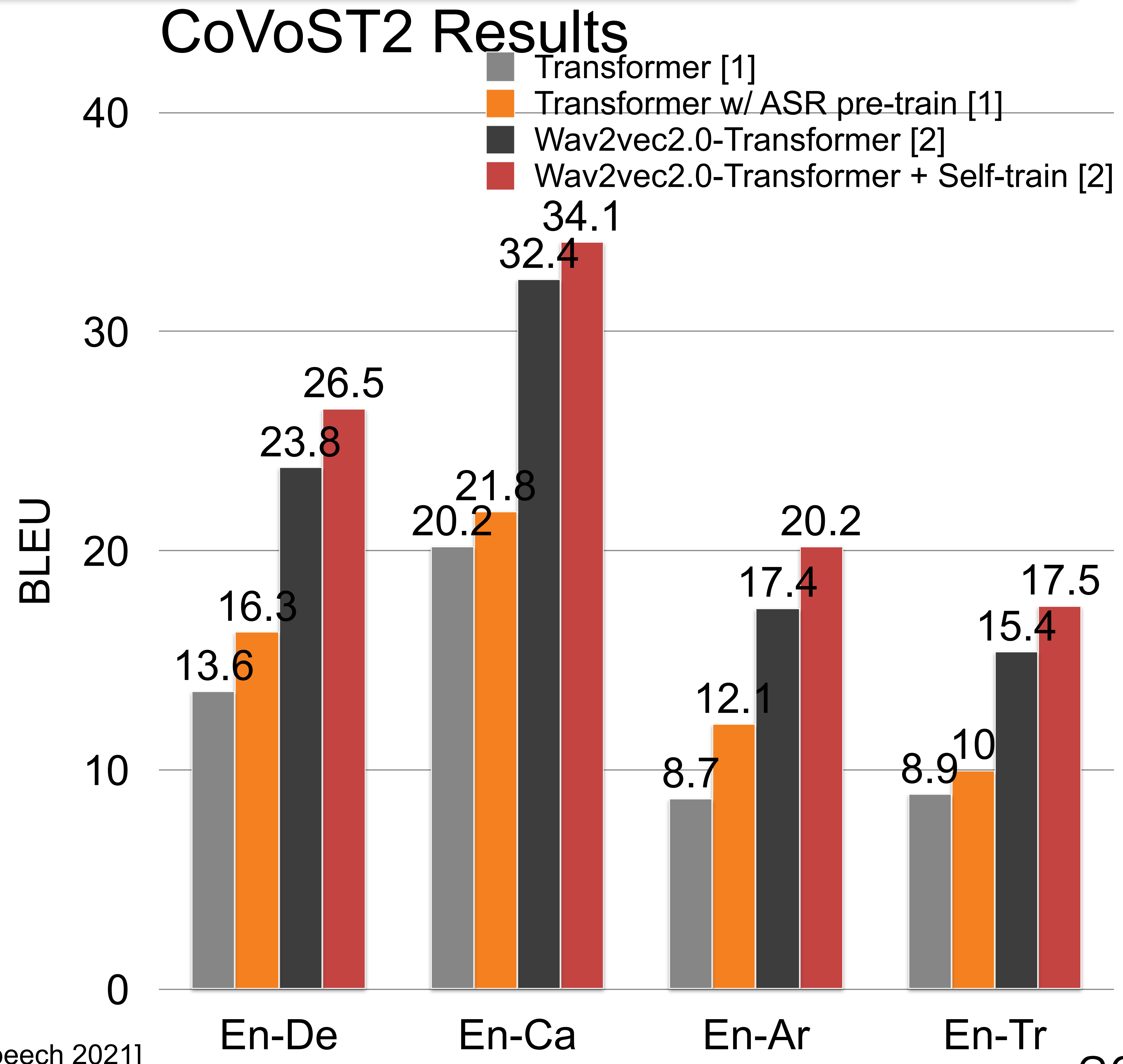
Self-training with Audio data



Step 0. Audio-only pre-training for Wav2vec2.0

Step 1. Freeze Wav2vec2.0, train on ST

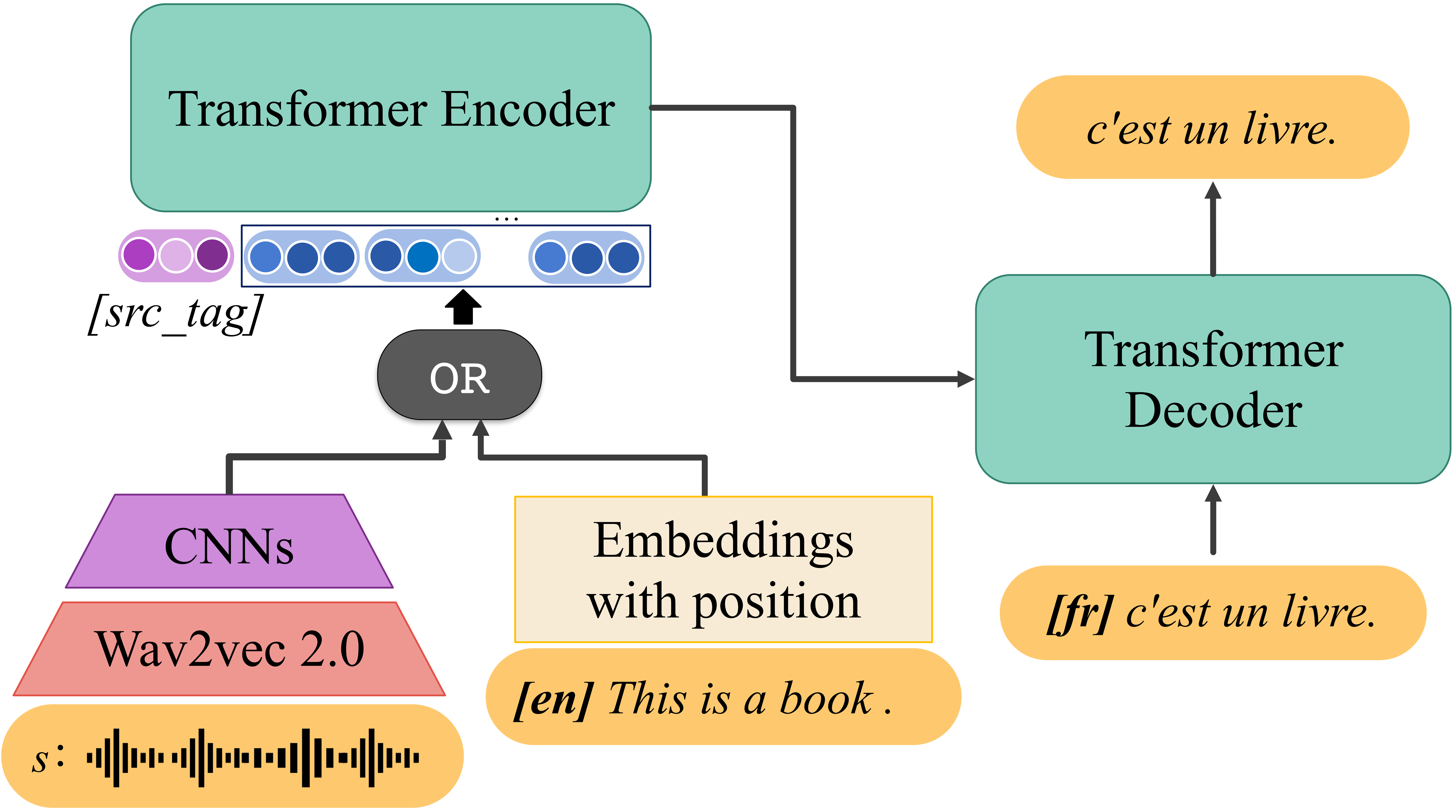
Step 2. Self-train on generated pseudo-translation with LibriVox audio



[1] CoVoST 2 and Massively Multilingual Speech-to-Text Translation, [Wang et al InterSpeech 2021]
[2] Large-Scale Self- and Semi-Supervised Learning for Speech Translation [Wang et al. 2021]

Fine-tuning Strategy for ST

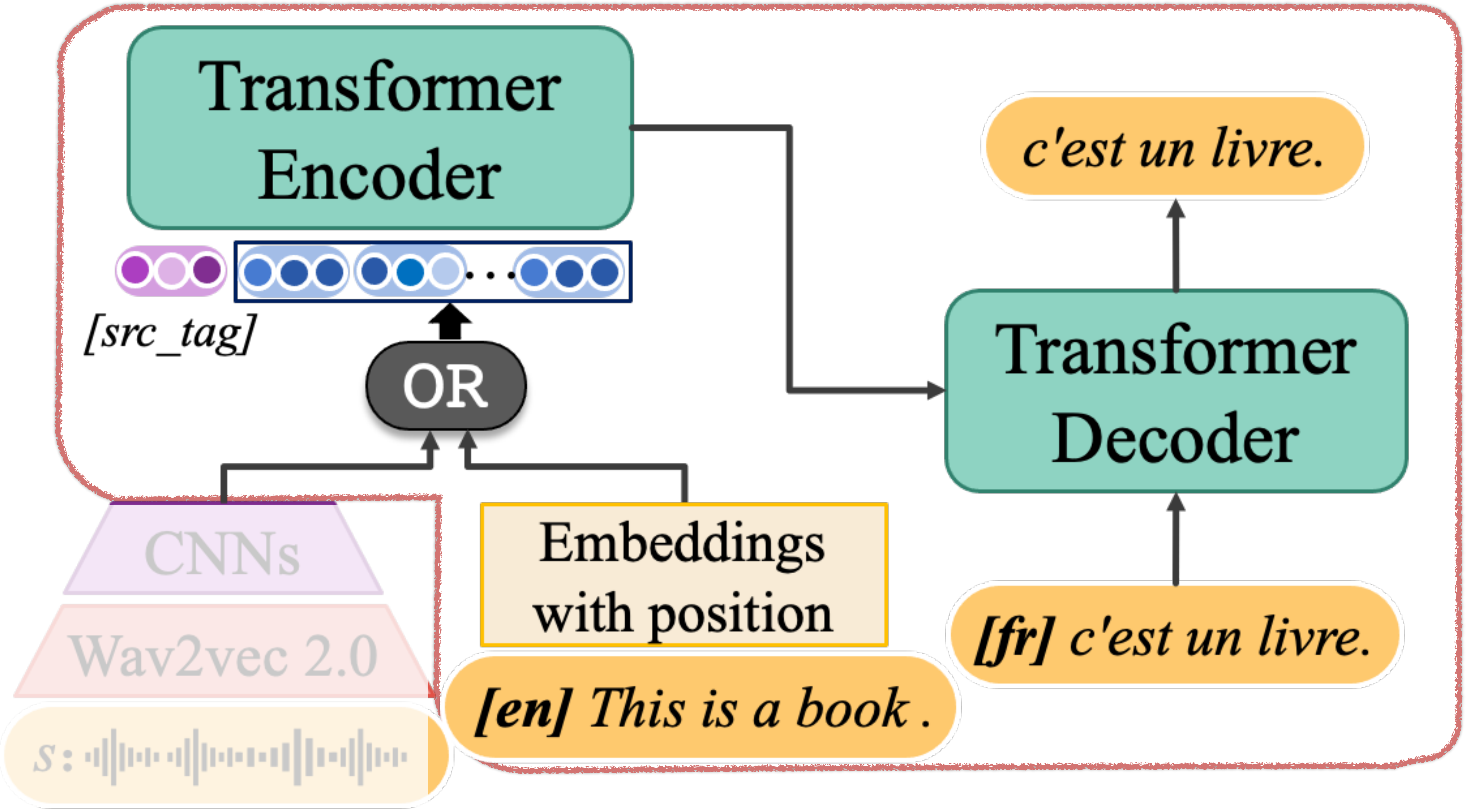
Cross Speech-Text Network (XSTNet)



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

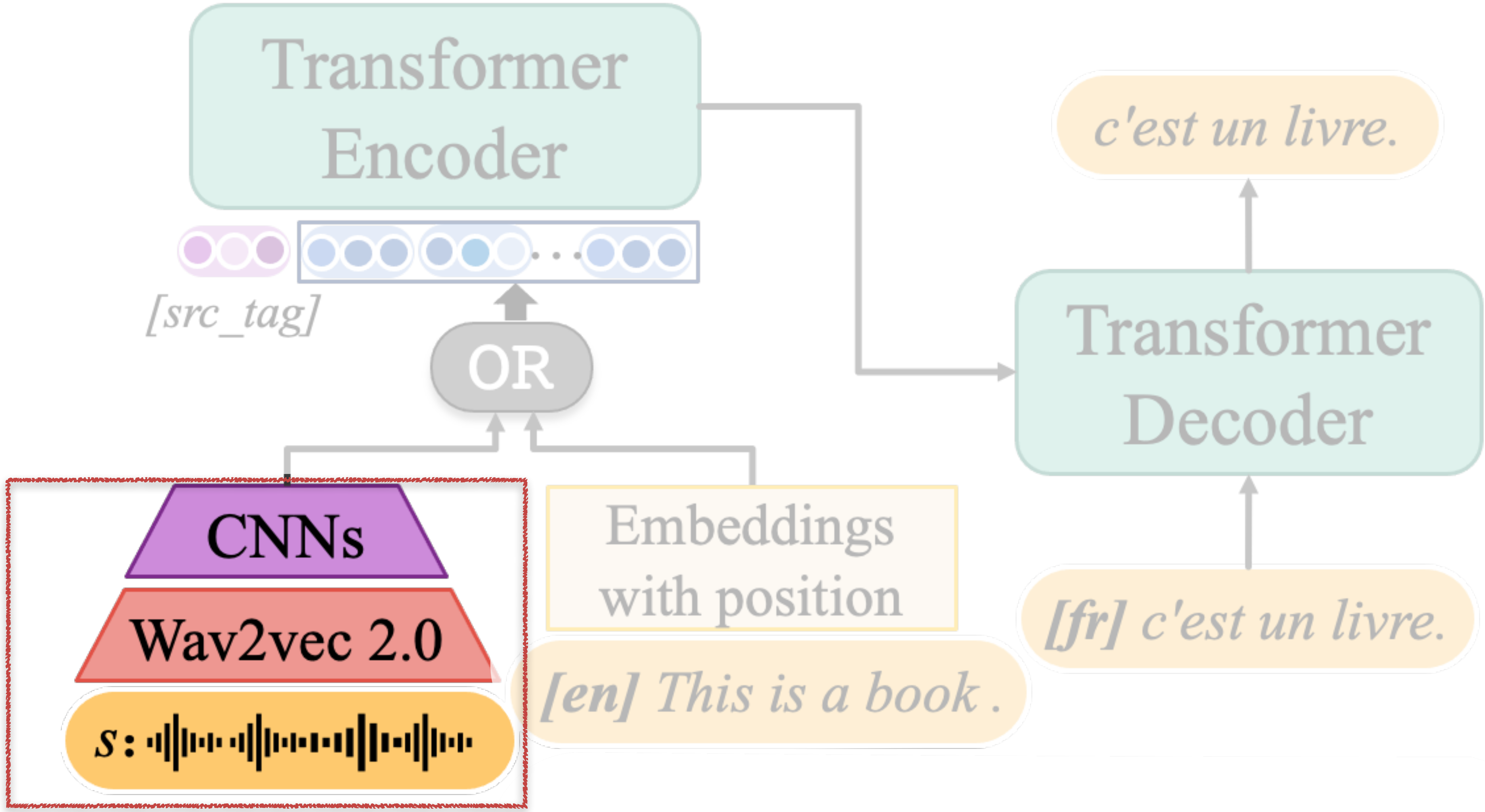


End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Supports inputs of two modalities

- ☑ Wav2vec2.0[1] as the acoustic encoder
- ☑ We add two convolution layers with 2-stride to shrink the length.

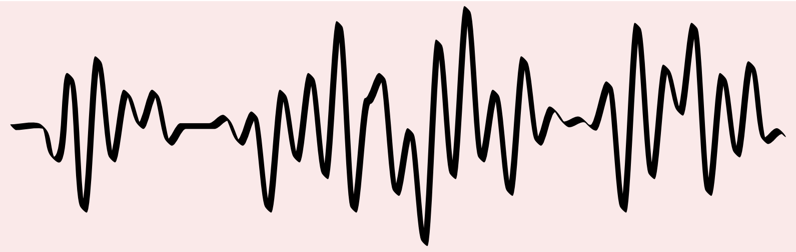
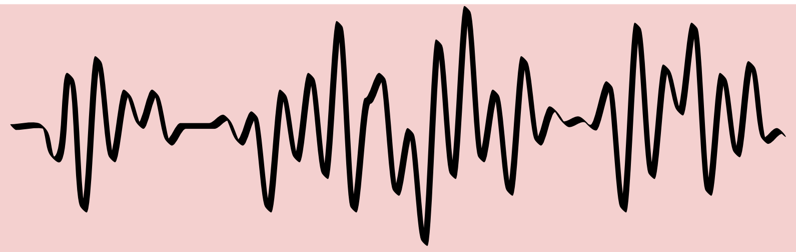
≈Acoustic encoder



[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}

Progressive:

Don't stop

training 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>*

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	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg.
Transformer ST [13]	×	ASR	22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
AFS [31]	×	×	22.4	26.9	31.6	23.0	24.9	26.3	21.0	14.7	23.9
Dual-Decoder Transf. [15]	×	×	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
Tang et al. [2]	MT	ASR, MT	23.9	28.6	33.1	-	-	-	-	-	-
FAT-ST (Big) [6]	ASR, MT, mono-data†	FAT-MLM	25.5	30.8	-	-	30.1	-	-	-	-
W-Transf.	audio-only*	SSL*	23.6	28.4	34.6	24.0	29.0	29.6	22.4	14.4	25.7
XSTNet (Base)	audio-only*	SSL*	25.5	29.6	36.0	25.5	30.0	31.3	25.1	16.9	27.5
XSTNet (Expand)	MT, audio-only*	SSL*, MT	**27.8**§	**30.8**	**38.0**	**26.4**	**31.2**	**32.4**	**25.7**	**18.5**	**28.8**

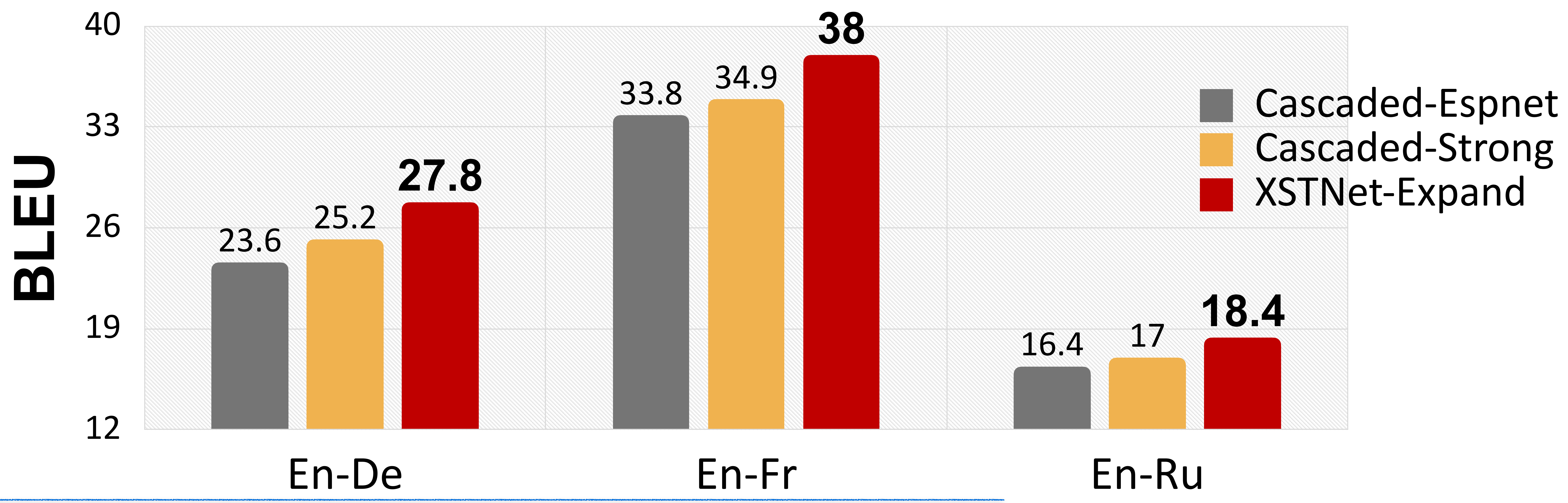
Table 1: *Performance (case-sensitive detokenized BLEU) on MuST-C test sets.* †: *"Mono-data" means audio-only data from Librispeech, Libri-Light, and text-only data from Europarl/Wiki Text;* *: *"Audio-only" data from LibriSpeech is used in the pre-training of wav2vec2.0-base module, and "SSL" means the self-supervised learning from unlabeled audio data.* § *uses OpenSubtitles as external MT 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

Learning Better Speech-Text Bimodal Representation

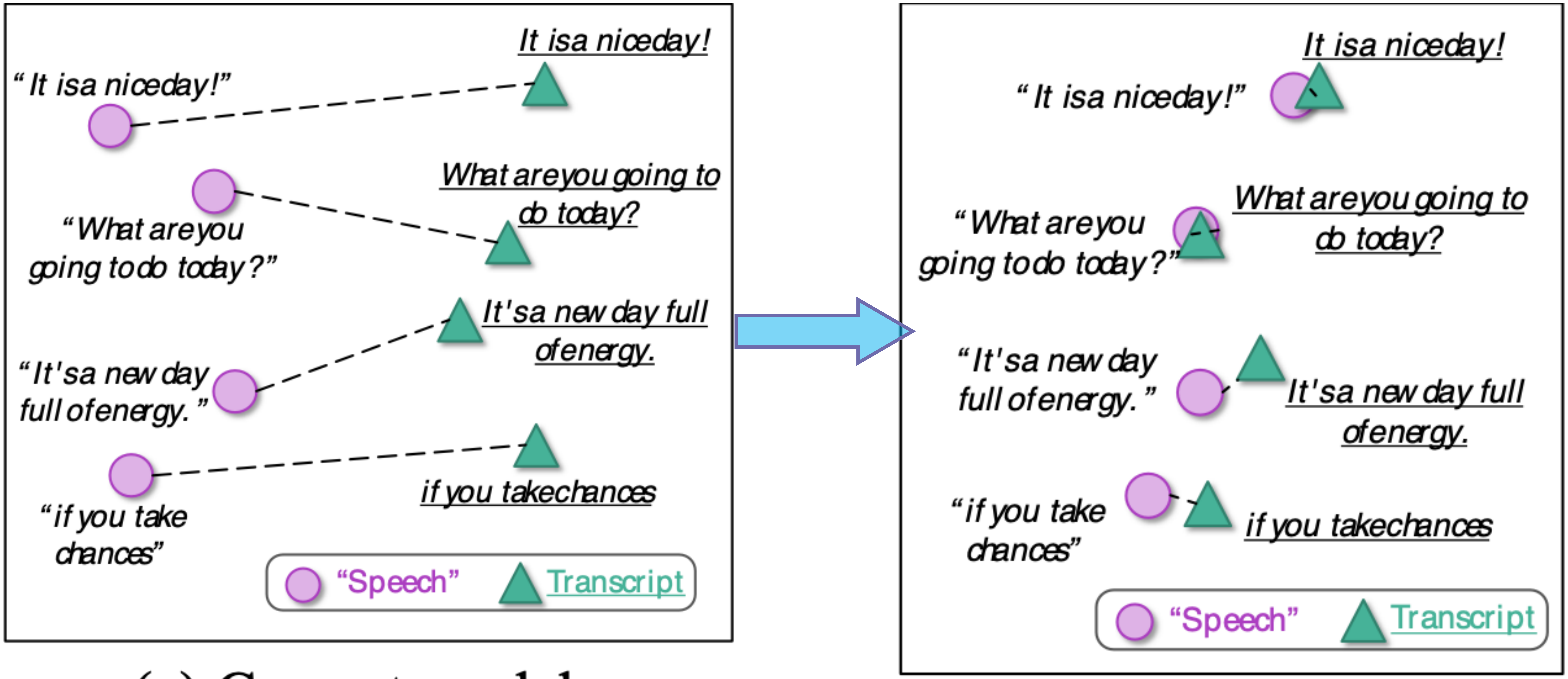
- ConST: Contrastive Learning to bridge the gap between text and speech [Ye et al 2022]
  - WACO: Contrastive learning at word-level with better aligned representation [Ouyang 2023]

- Chimera: Learning Fixed-size Shared Space for both audio and text, audio+MT pretraining [Han et al. 2021]

- Wav2vec2.0-mTransformer LNA: Use both audio pertaining + multilingual pertained language model, and selective efficient fine-tuning [Li et al. ACL 2021]

- FAT-ST: Masked pre-training for fused audio and text [Zheng et al. ICML 2021]

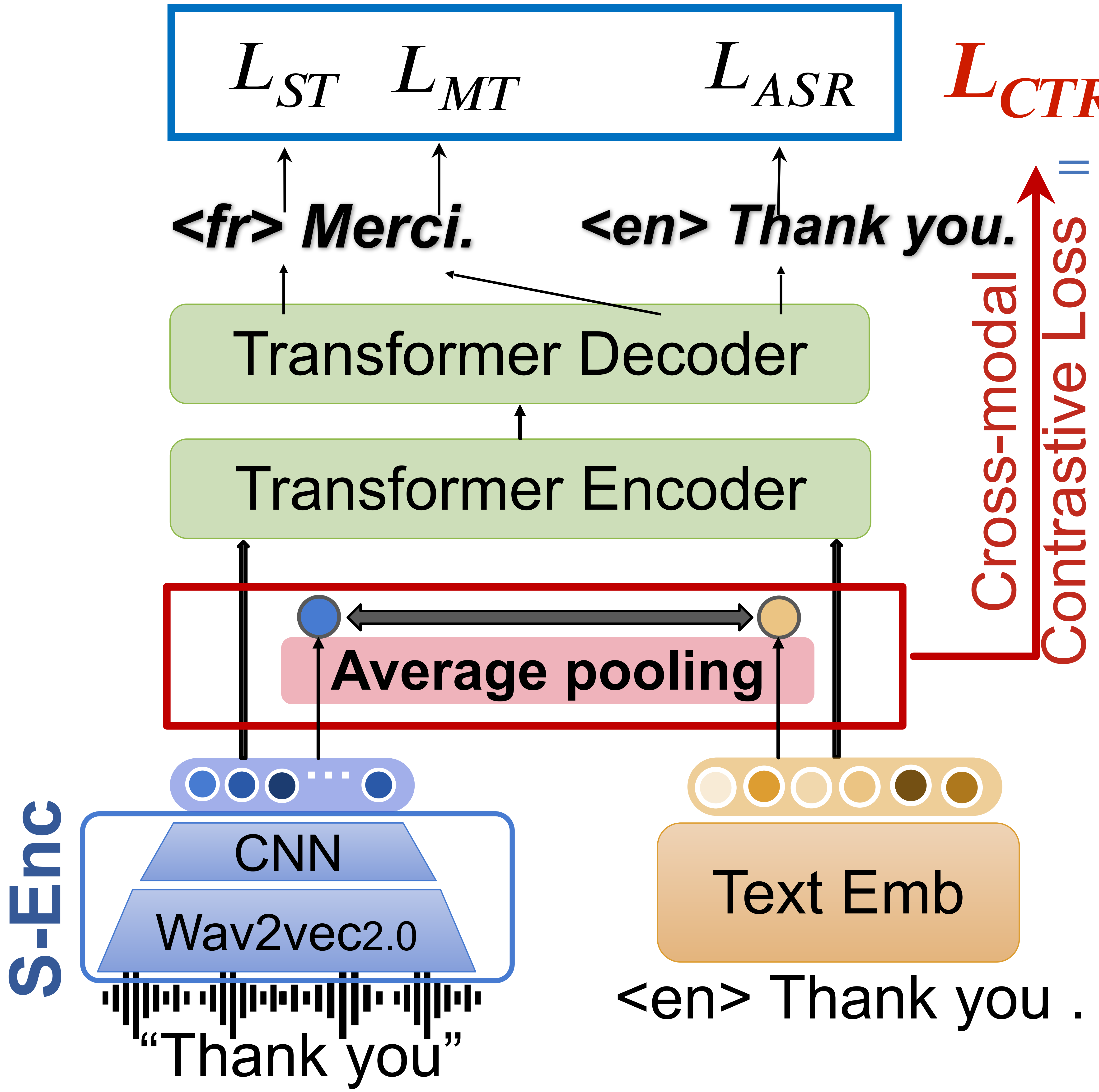
Text and speech with same meaning should be **similar** in representation!



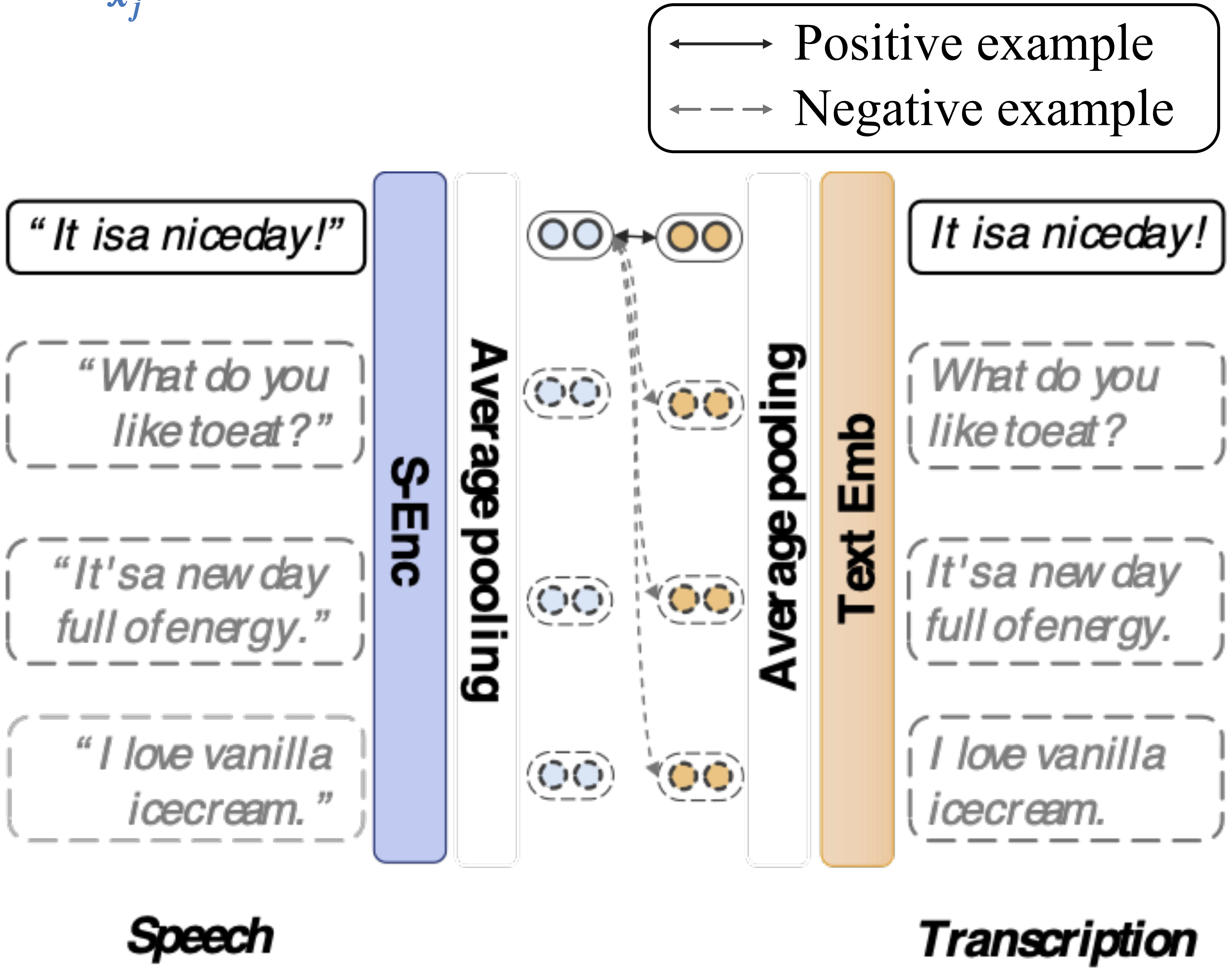
(a) Current models

(b) Expected

Contrastive Learning for Speech Translation

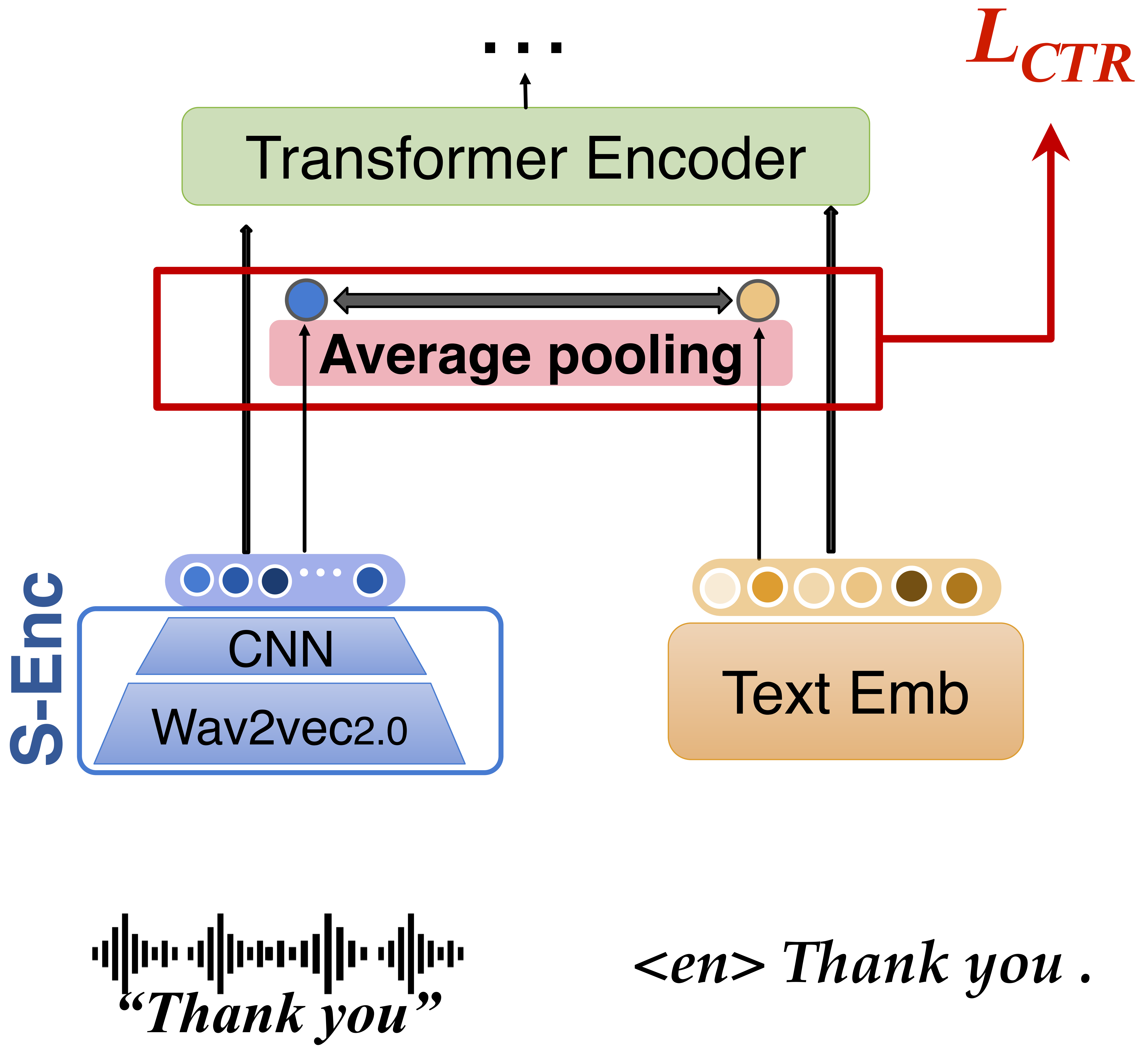


$$L_{CTR} = -\sum_{s,x} \log \frac{e^{\cos(u,v)/\tau}}{\sum_{x_j} e^{\cos(u,v_j)/\tau}}$$



Cross-modal Contrastive Learning for Speech Translation,[Ye, Wang, **Lei Li**, NAACL 2022]

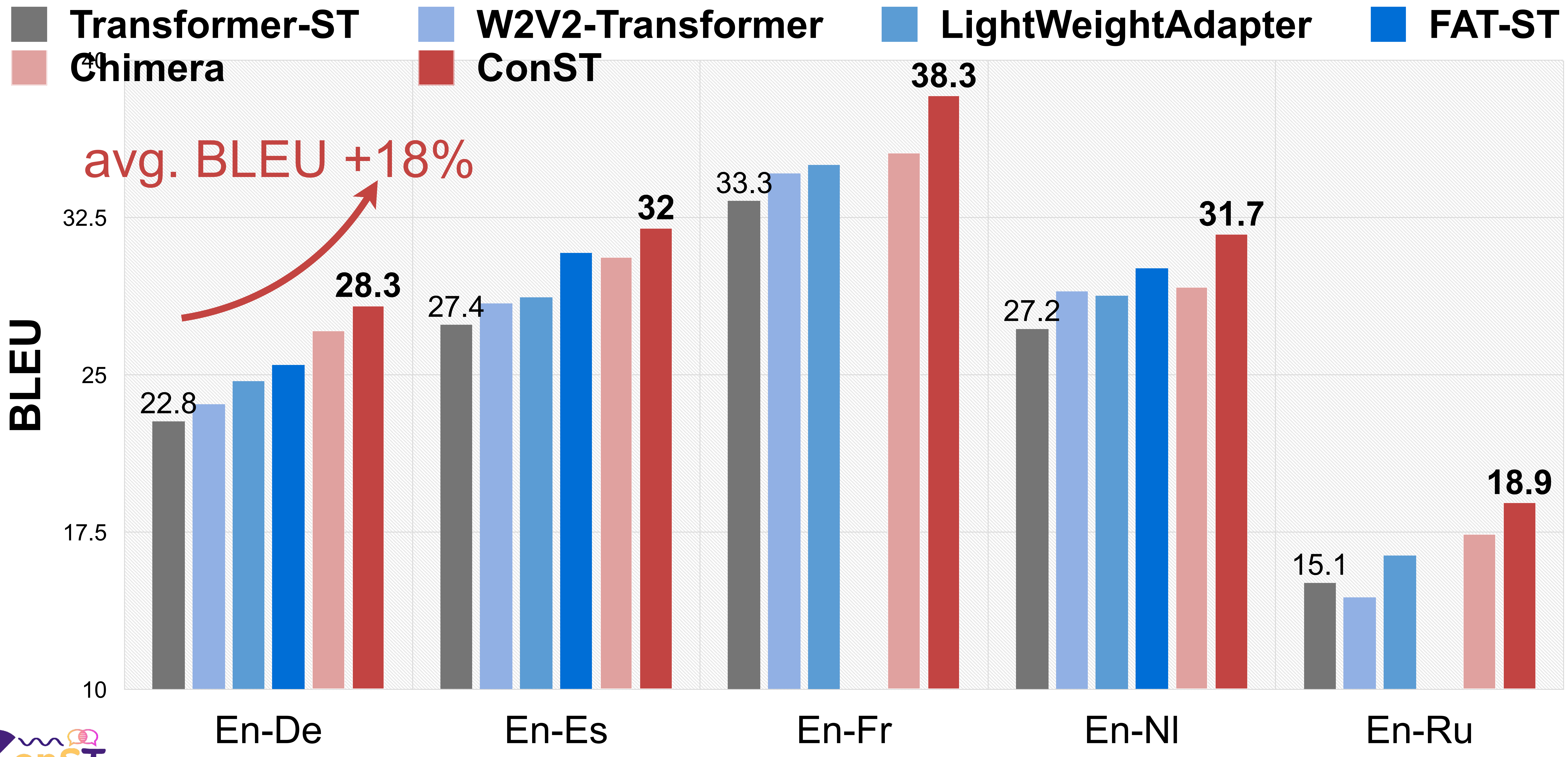
Mining more hard examples



We introduce three hard example mining operations.

1. Span-Masked Aug. (SMA)
2. Word Repetition (Rep)
3. Cut-off

Proposed ConST Significantly Improves Translation Performance

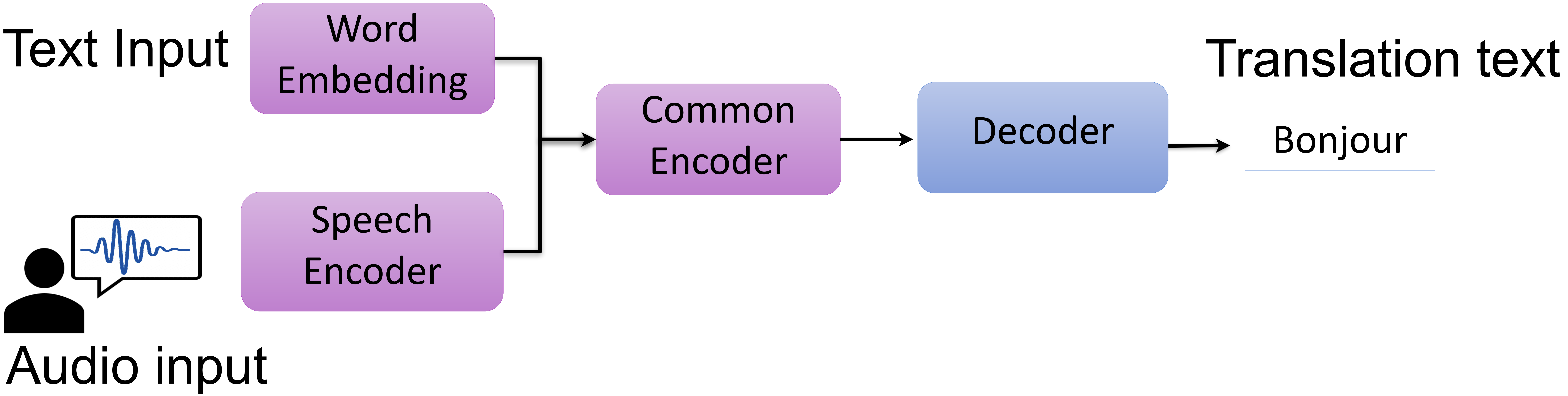


Cross-modal Contrastive Learning for Speech Translation,[Ye, Wang, **Lei Li**, NAACL 2022]

Both Multi-task and Contrastive Learning are important!

$$\mathcal{L} = \boxed{\mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{ASR}} + \mathcal{L}_{\text{MT}}} + \lambda\mathcal{L}_{\text{CTR}}$$

Bi-modal Encoding Architecture for ST

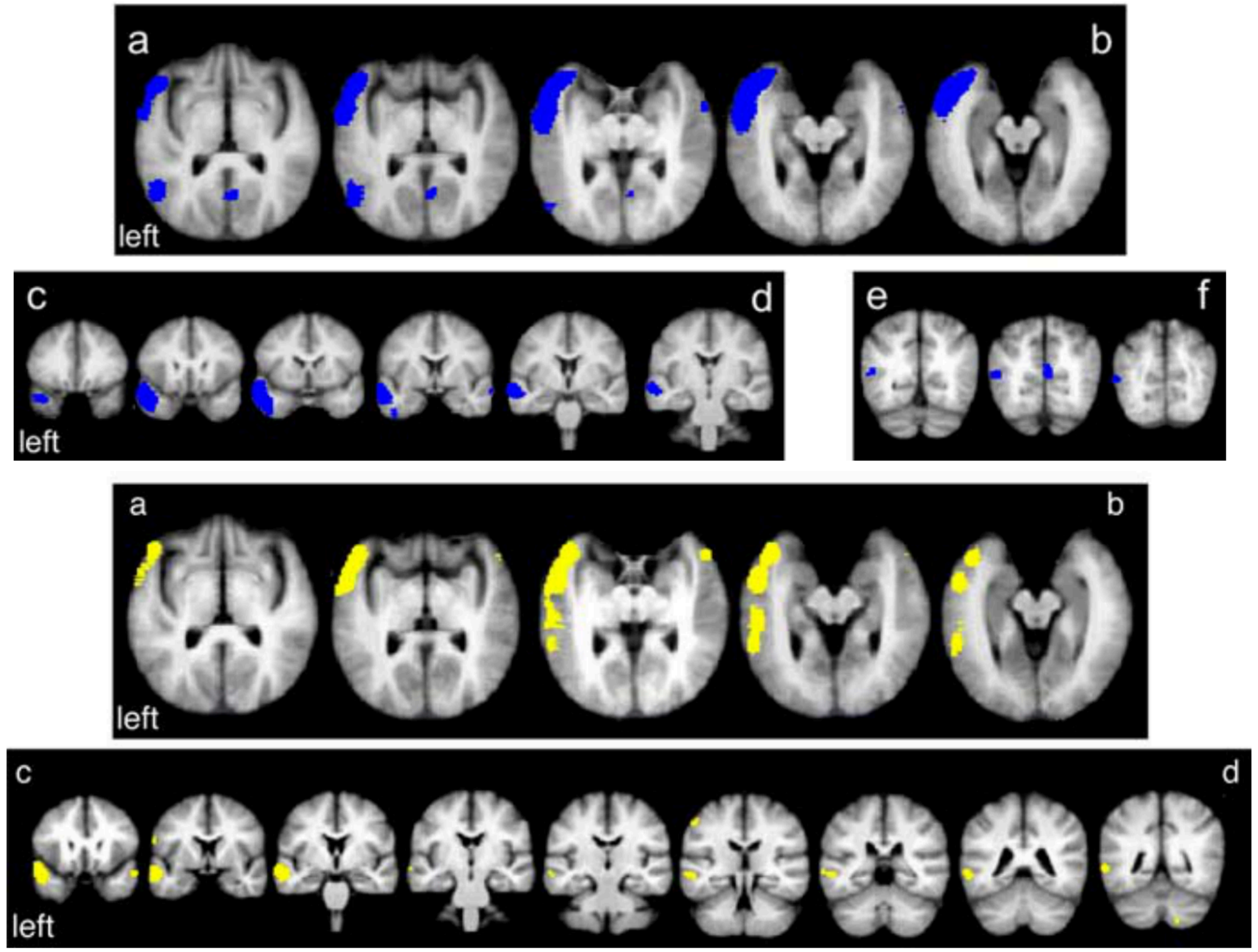


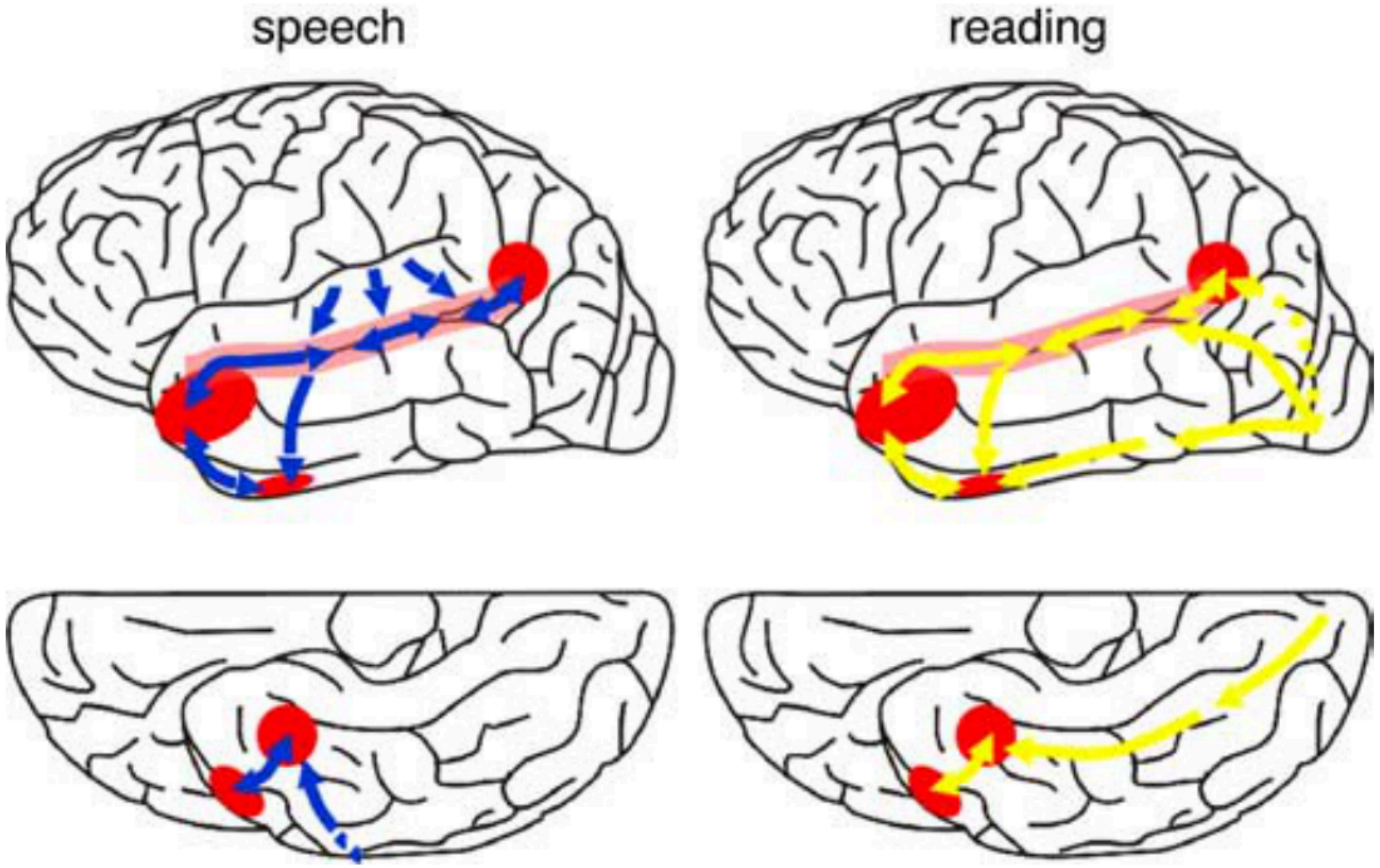
Challenges: gap between text and audio

1. Length: ~20 (text) vs. ~ 1k-10k (audio)
2. Embedding space disparity

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)

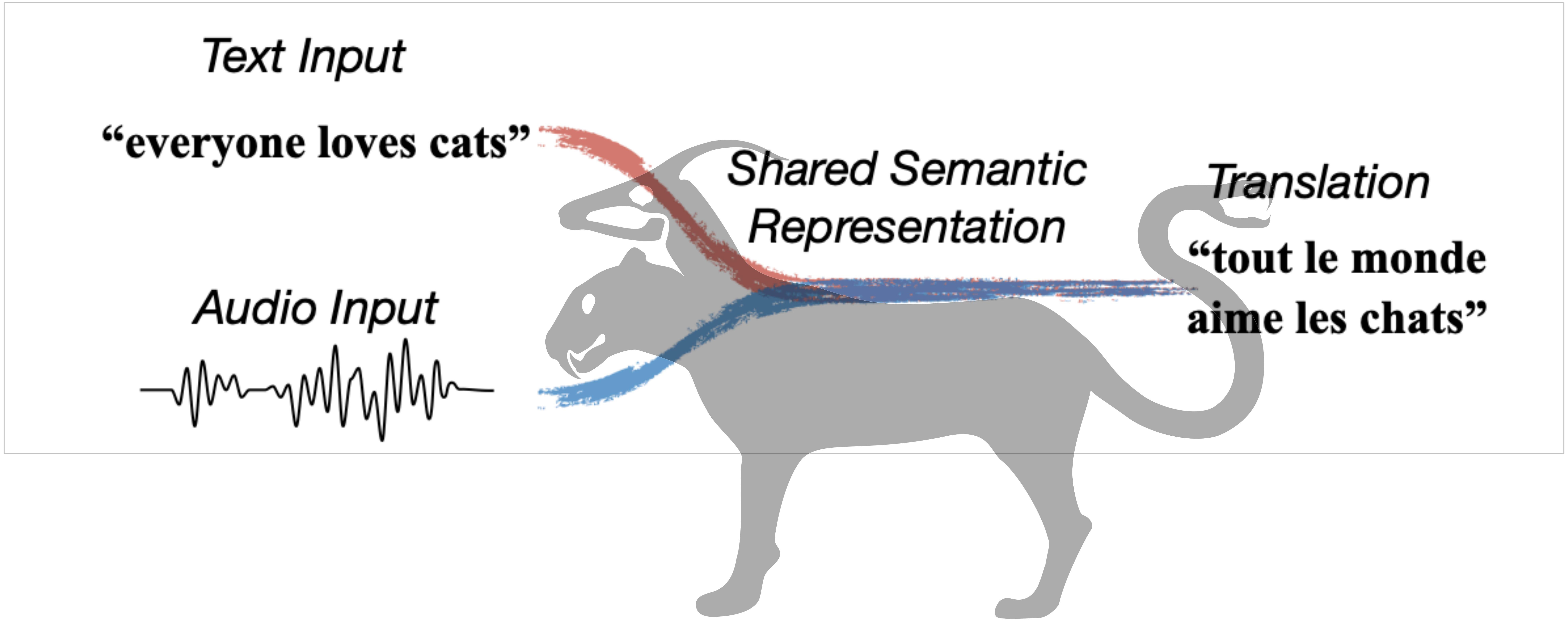
[1] Van Atteveldt, Nienke, et al. "Integration of letters and speech sounds in the human brain." *Neuron* 43.2 (2004): 271-282.

[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>

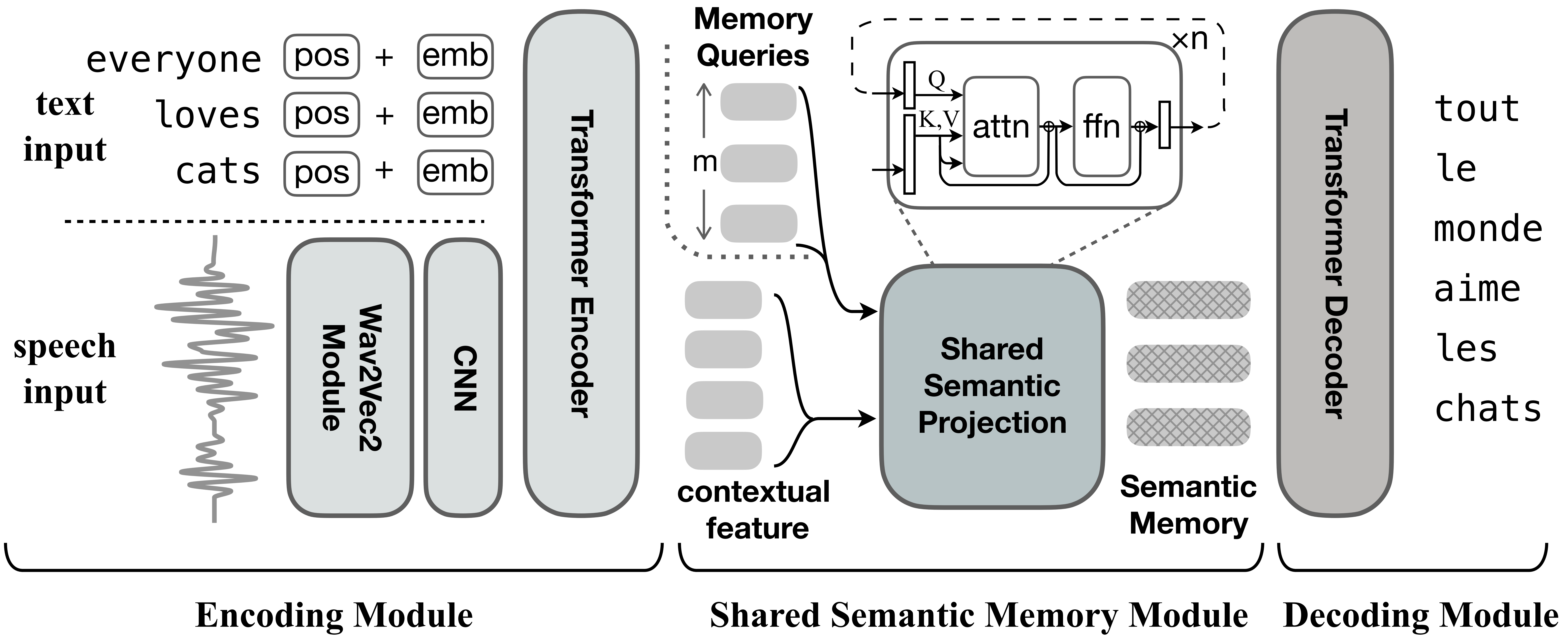


Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Chimera Model for ST

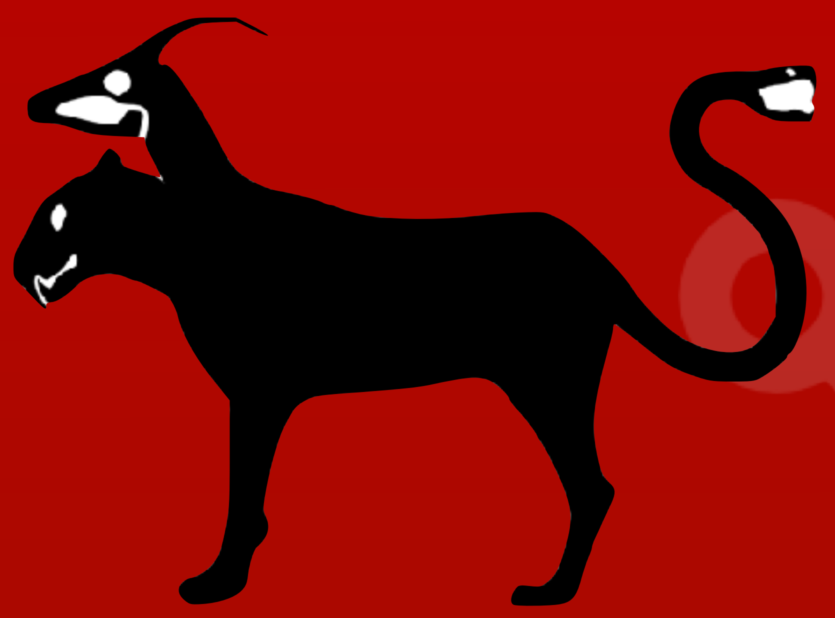
Training with auxiliary objectives: ST + MT + Contrastive loss

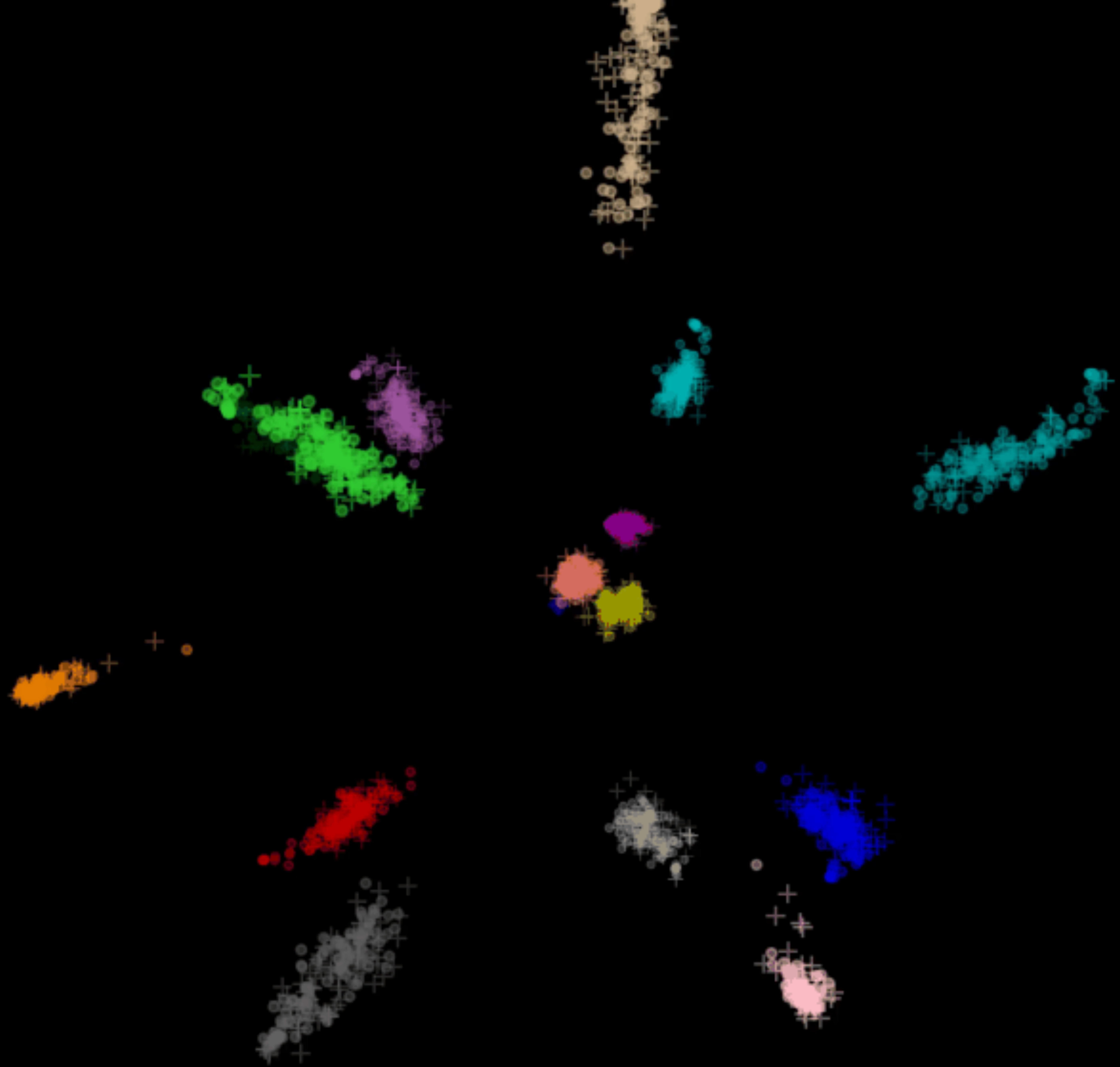
Benefit: able to exploit large external MT data



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

Shared Semantic Space Learned by Proposed Chimera



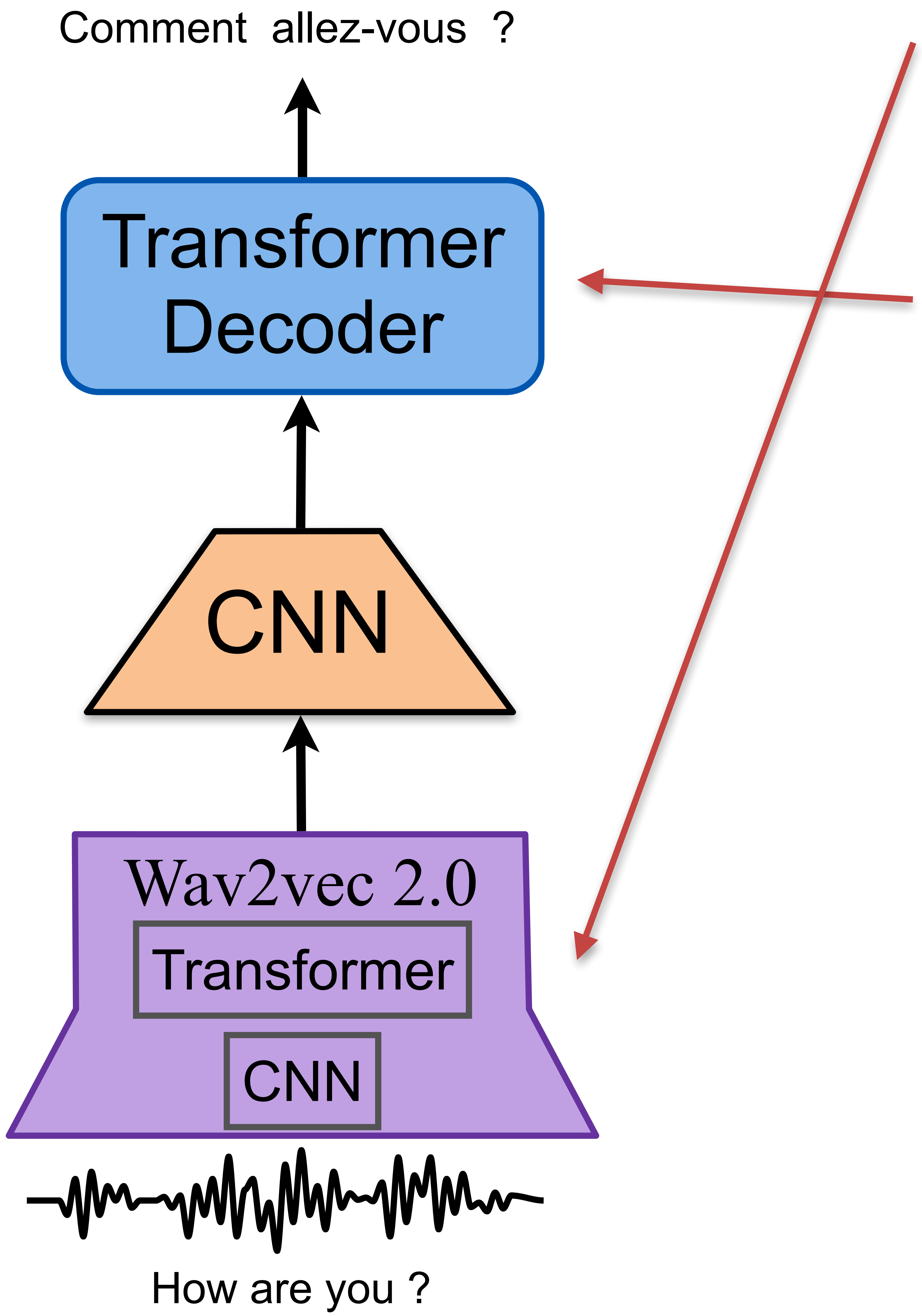


Chimera achieves the best (so far) BLEU on all languages in MuST-C

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 [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Audio and Multilingual Text Pretrain for Multilingual ST



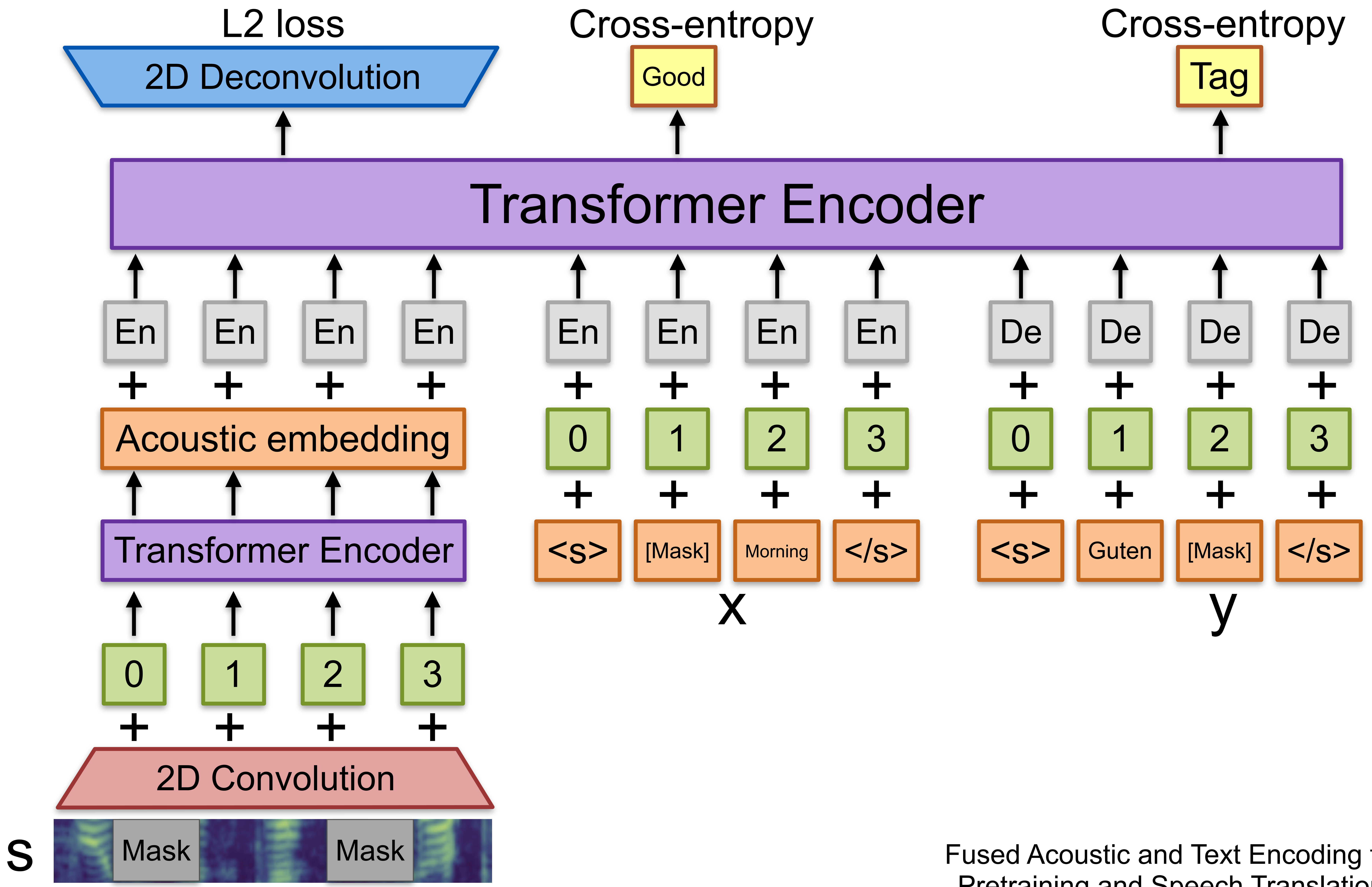
- Encoder uses Wav2vec2.0 pre-trained on LibriVox-60k audio
- Decoder: mBart pre-trained on 50 monolingual text and 49 bitext
- ST finetune strategy (LNA):
  - Only fine-tune layer-norm and attention layers
- MT+ST multitask joint train with further parallel bitext data

Multilingual Speech Translation with Efficient Finetuning of Pretrained Models [Li et al, ACL 2021]

Wav2vec2.0 retraining + Multilingual training effectively transfers to low resource source language

Multilingual Speech Translation with Efficient Finetuning of Pretrained Models [Li et al, ACL 2021]

Fused Acoustic and Text Masked Language Model (FAT-MLM)

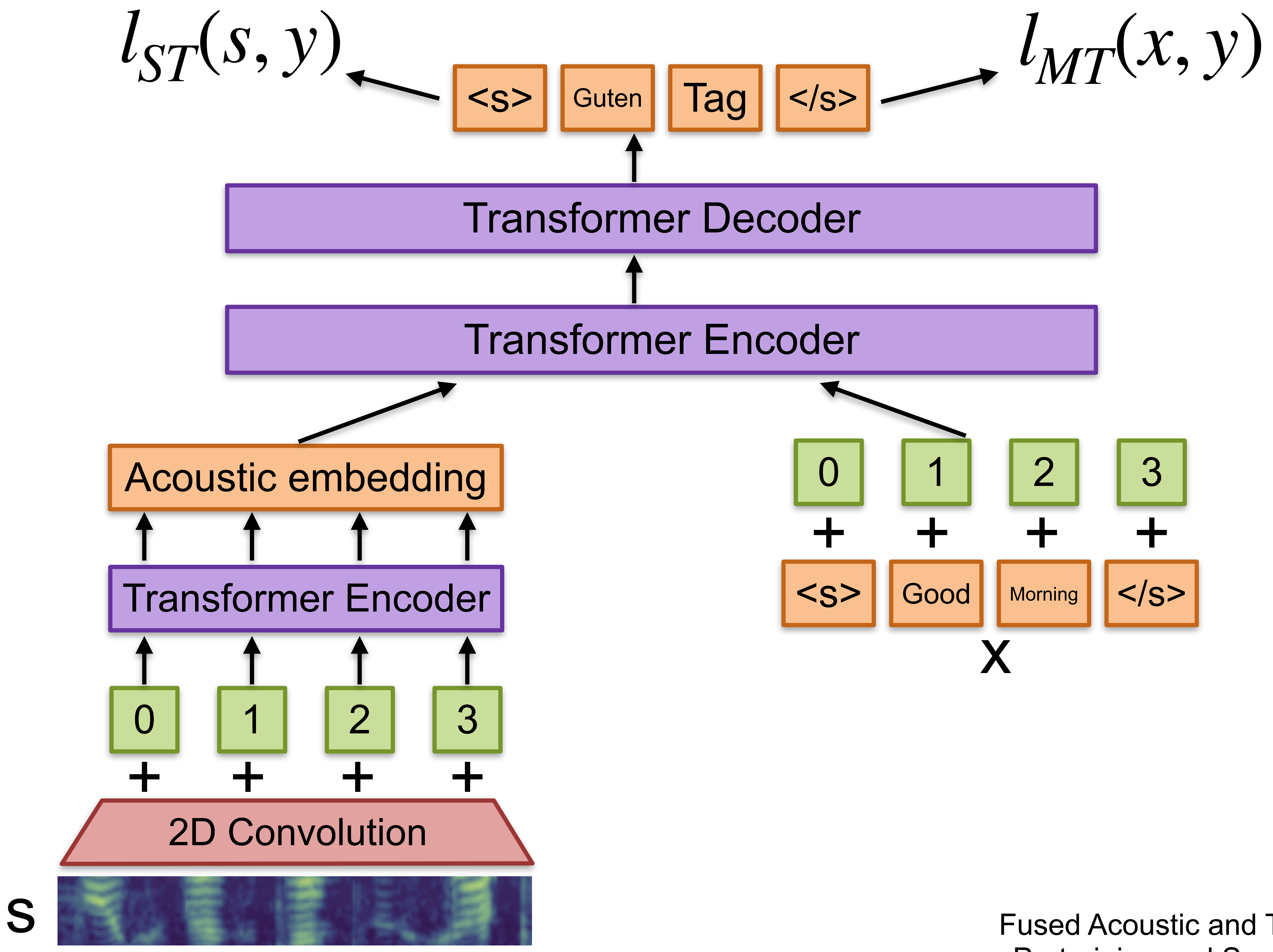


Pre-training data

1. Librispeech ASR 960h
2. Libri-light audio 3,748h
3. Europarl/wiki text 2.3M
4. MuST-C 408h
5. Europarl MT 1.9M

Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

FAT-ST



Training:

- Pre-train FAT-MLM with all data
- Init FAT-ST with FAT-MLM, decoder copy encoder
- Further fine-tune on MuST-C ST data.

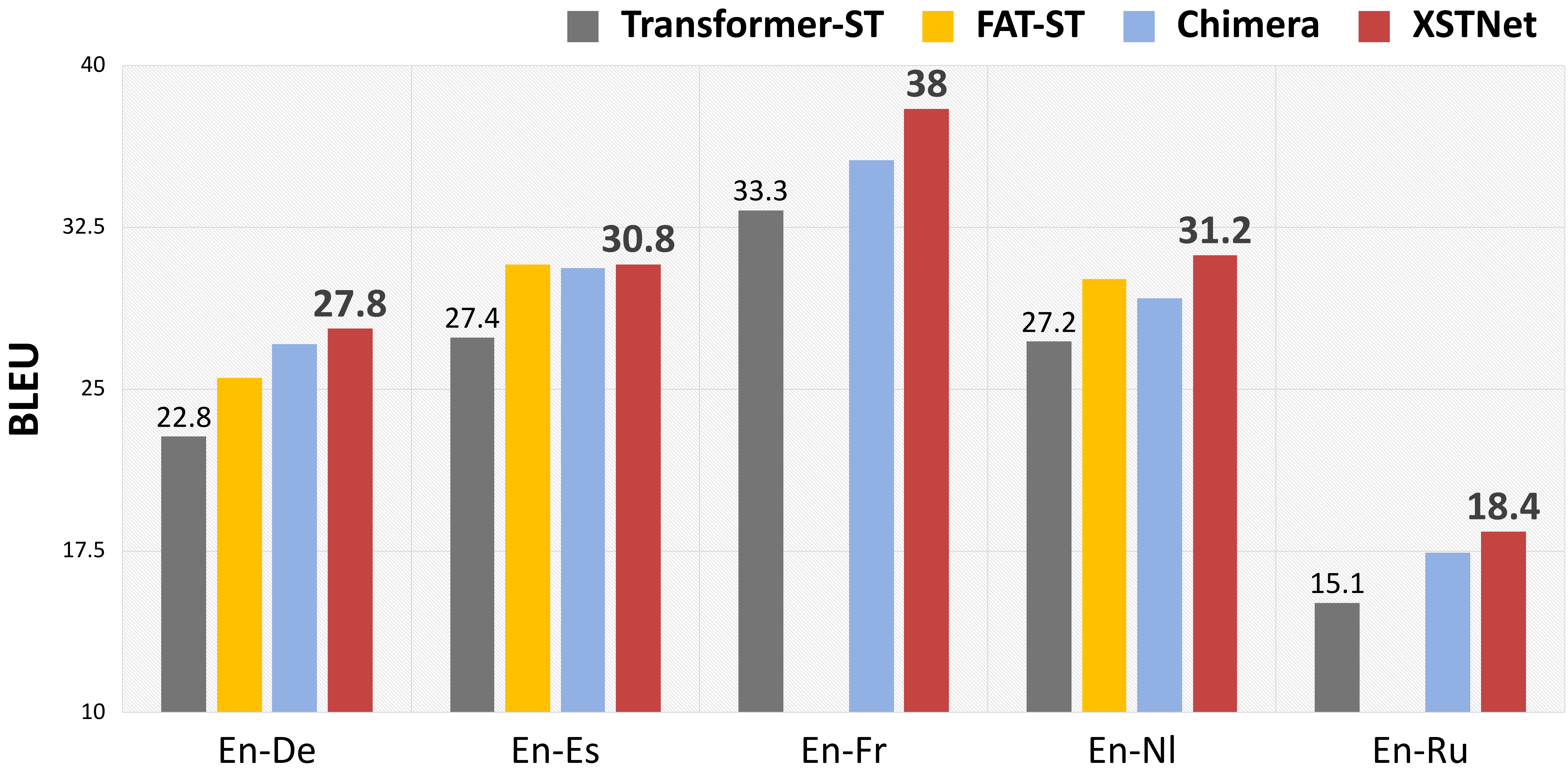
Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

Joint audio&text Pre-training task helps ST

Pretrain Method	Models	En→De	En→Es	En→Nl	Avg.	Model Size
No Pretraining	ST	19.64	23.68	23.01	22.11	31.25M
	ST + ASR	21.70	26.83	25.44	24.66 (+2.55)	44.82M
	ST + ASR & MT	21.58	26.37	26.17	24.71 (+2.60)	56.81M
	ST + MAM	20.78	25.34	24.46	23.53 (+1.42)	33.15M
	ST + MAM + ASR	22.41	26.89	26.49	25.26 (+3.15)	46.72M
	Liu et al. (2020b)	22.55	-	-	-	-
	Le et al. (2020)	23.63	28.12	27.55	26.43 (+4.32)	51.20M
	Cascade[§]	23.65	28.68	27.91	26.75 (+4.64)	83.79M
	FAT-ST (base).	22.70	27.86	27.03	25.86 (+3.75)	39.34M
ASR & MT	ST	21.95	26.83	26.03	24.94 (+2.83)	31.25M
	ST + ASR & MT	22.05	26.95	26.15	25.05 (+2.94)	56.81M
MAM	FAT-ST (base)	22.29	27.21	26.26	25.25 (+3.14)	39.34M
FAT-MLM	FAT-ST (base)	**23.68**	28.61	**27.84**	26.71 (+4.60)	39.34M
	FAT-ST (big)	23.64	**29.00**	27.64	**26.76** (+4.65)	58.25M

Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

Pre-training Improves ST Performance



Summary

	Direct Supervision	Contrastive	Masked LM	Knowledge distillation	Progressive train	Selective Fine-tune	Self-training
MT Parallel Text	COSTT			[Liu et al. 2019]	XSTNet		
ASR Speech-Transcript	LUT	ConST, WACO					
Audio-only		Wav2vec Wav2vec 2.0					[Wang et al. 2021]
Raw text				LUT			
Speech+Text		Chimera	FAT-ST		XSTNet	LNA	

Language in 10