CS11-737 Multilingual NLP

Streaming Speech Translation

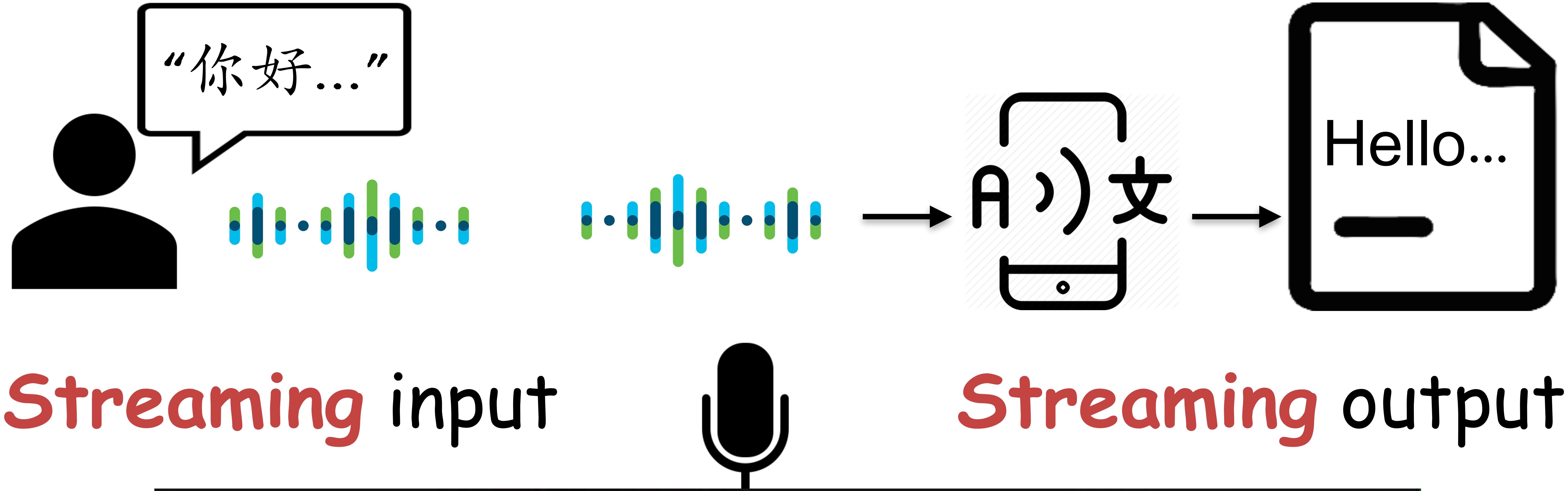
Lei Li

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

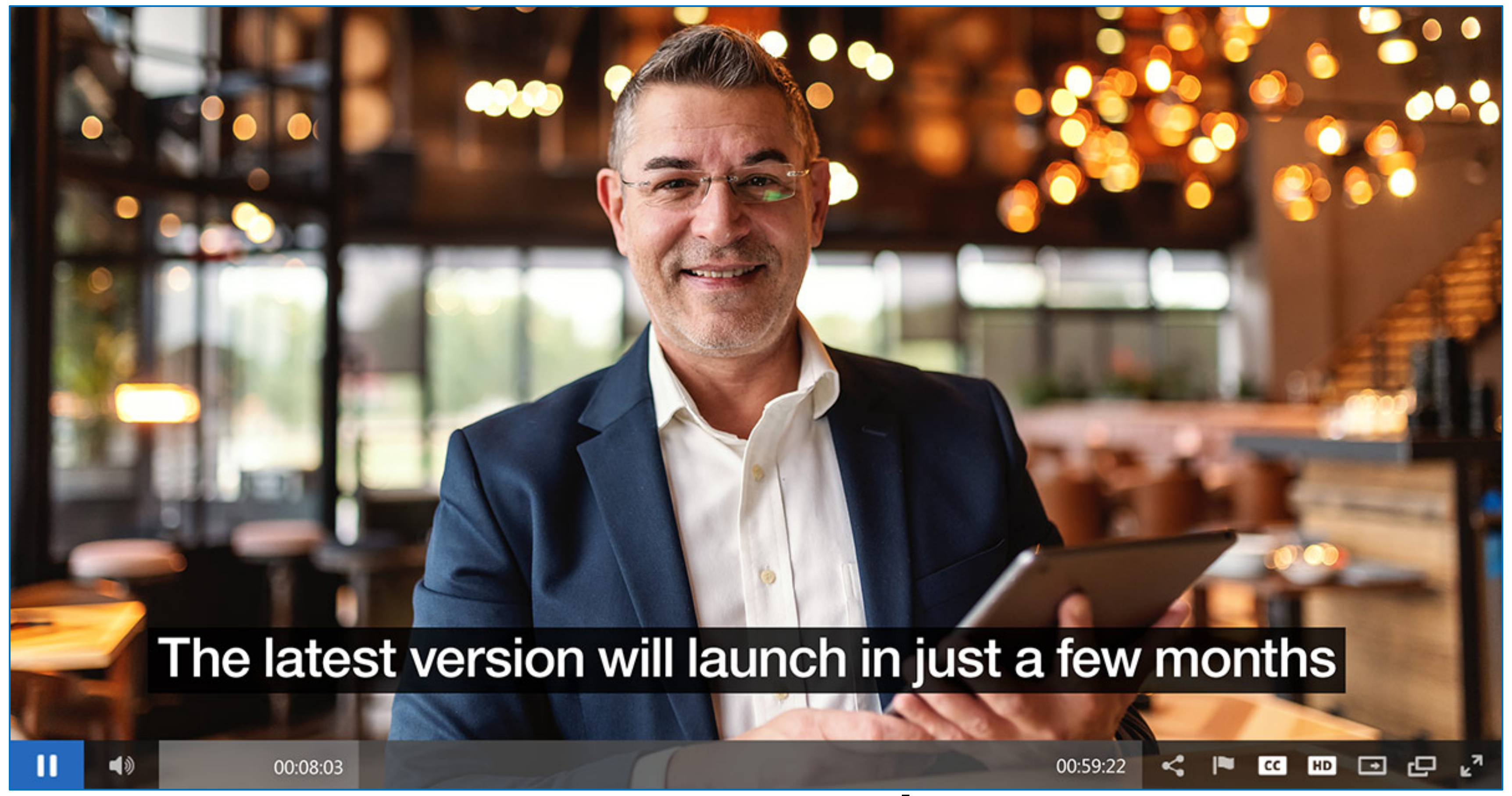
Carnegie Mellon University
Language Technologies Institute

Simultaneous Speech-to-text Translation

- Read the audio signals of speech in one language, and translate to the text in another language while speaker speaks (SiST).



Wide Applications of SST



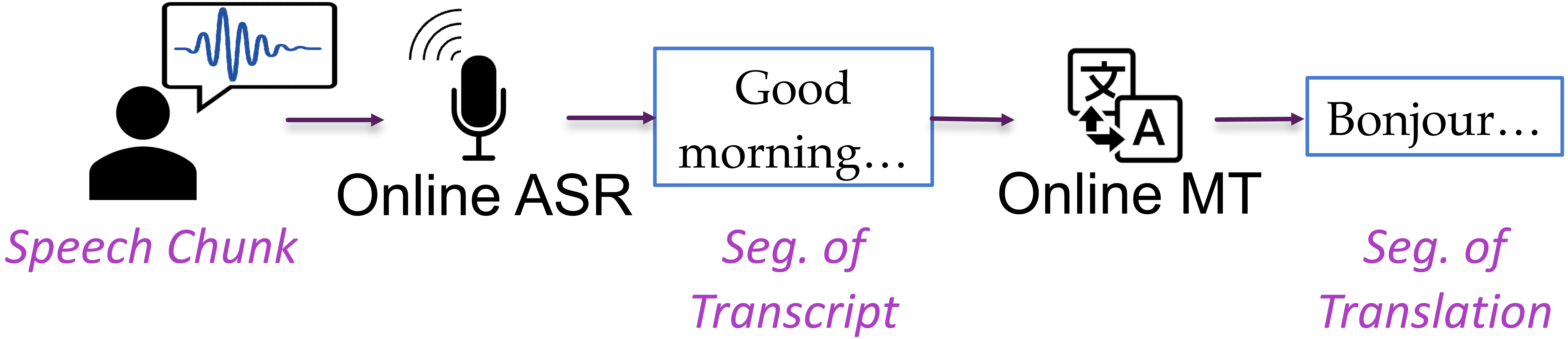
Foreign Media


Global Conferences

Tourism


International Trade

Traditional Cascaded SiST System



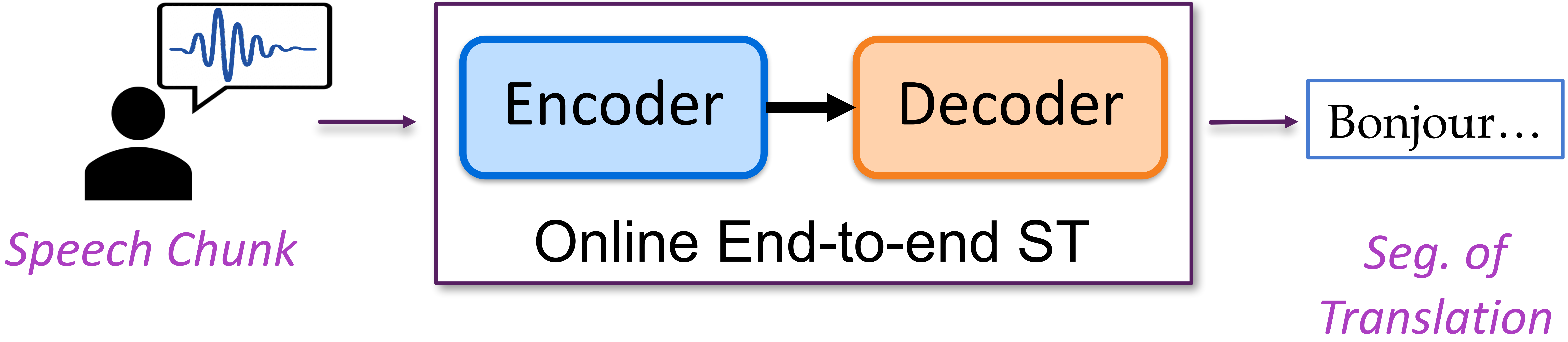
- Drawbacks:

1. Computationally inefficient

2. Error propagation:

Wrong/error transcript recognition → Wrong translation

End-to-end SiST



- **Goal:** End2end streaming ST needs to balance the latency and quality, and generate translations based on the partial speech chunk with a single model.
- **Predecessor's method:** Wait-K

Challenges for SiST

- Latency
- Applicability
- ...

Low latency is required for better user experience. —> Translate as early as possible.

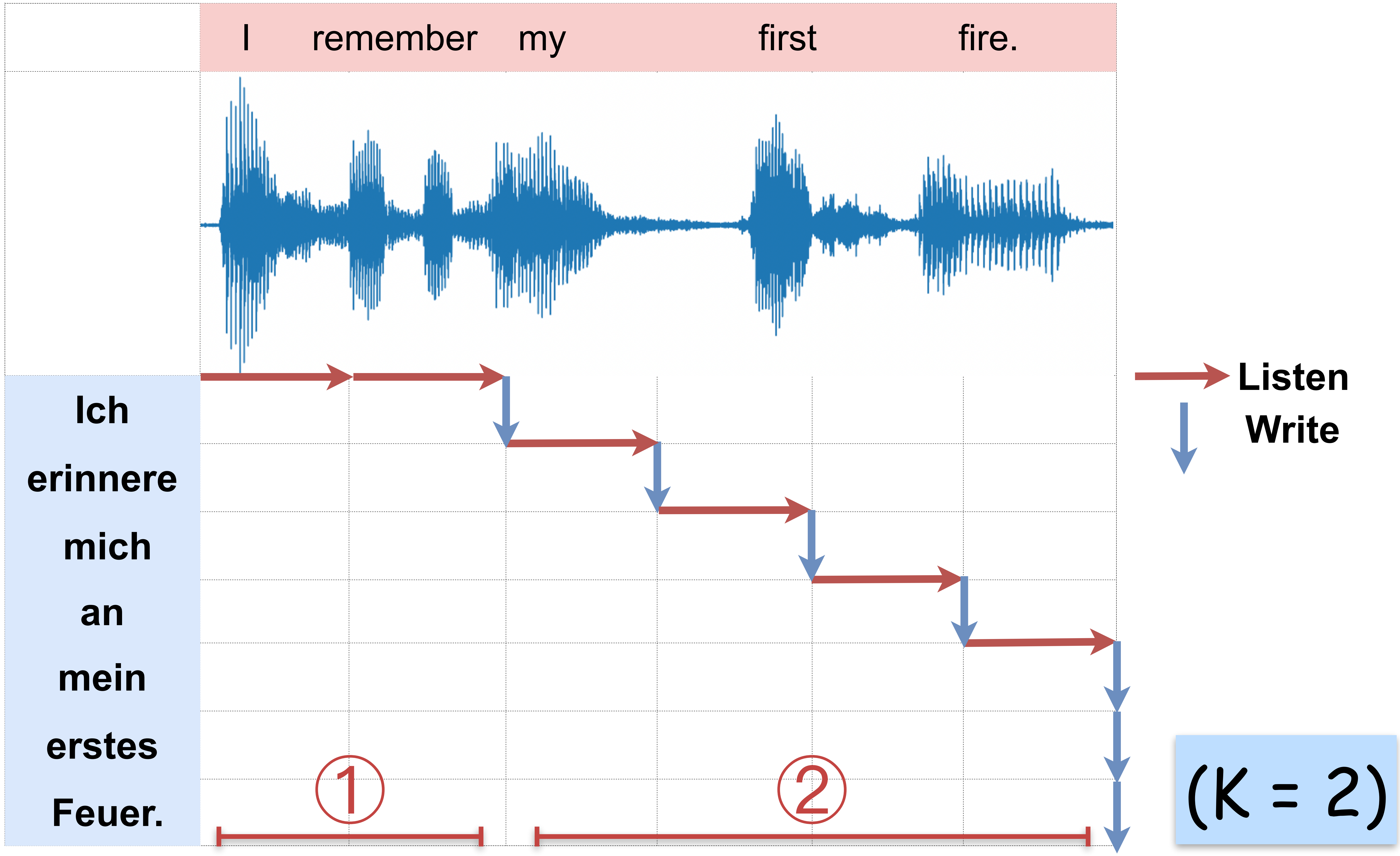
- Accuracy
- Flicker
- ...

More context is required to improve speech translation. —> Wait as long as possible.

Simple Approach: Wait-K with Fixed Stride

① Listen to streaming speech with **a fixed stride** after **K** steps.

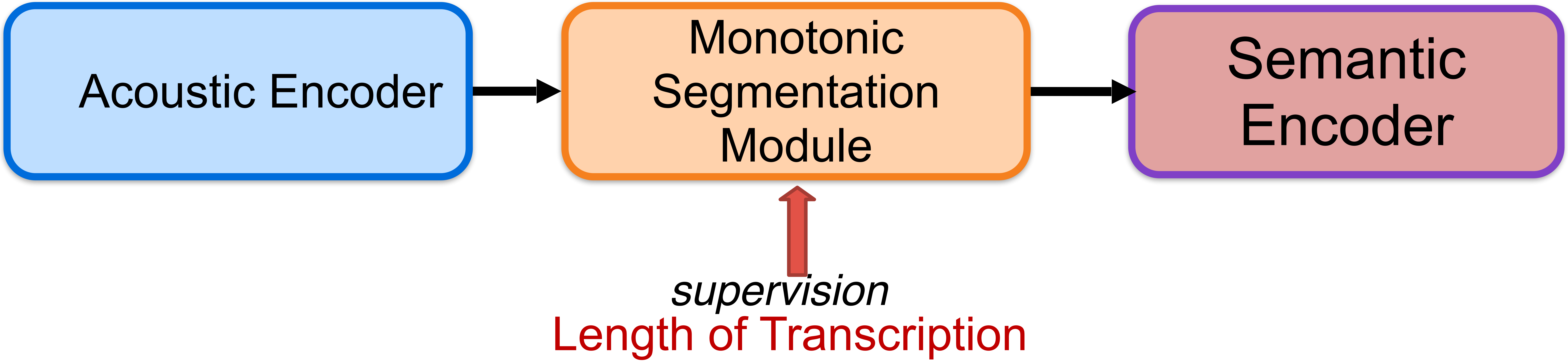
② Do **listen** and **write** iteratively till the end.



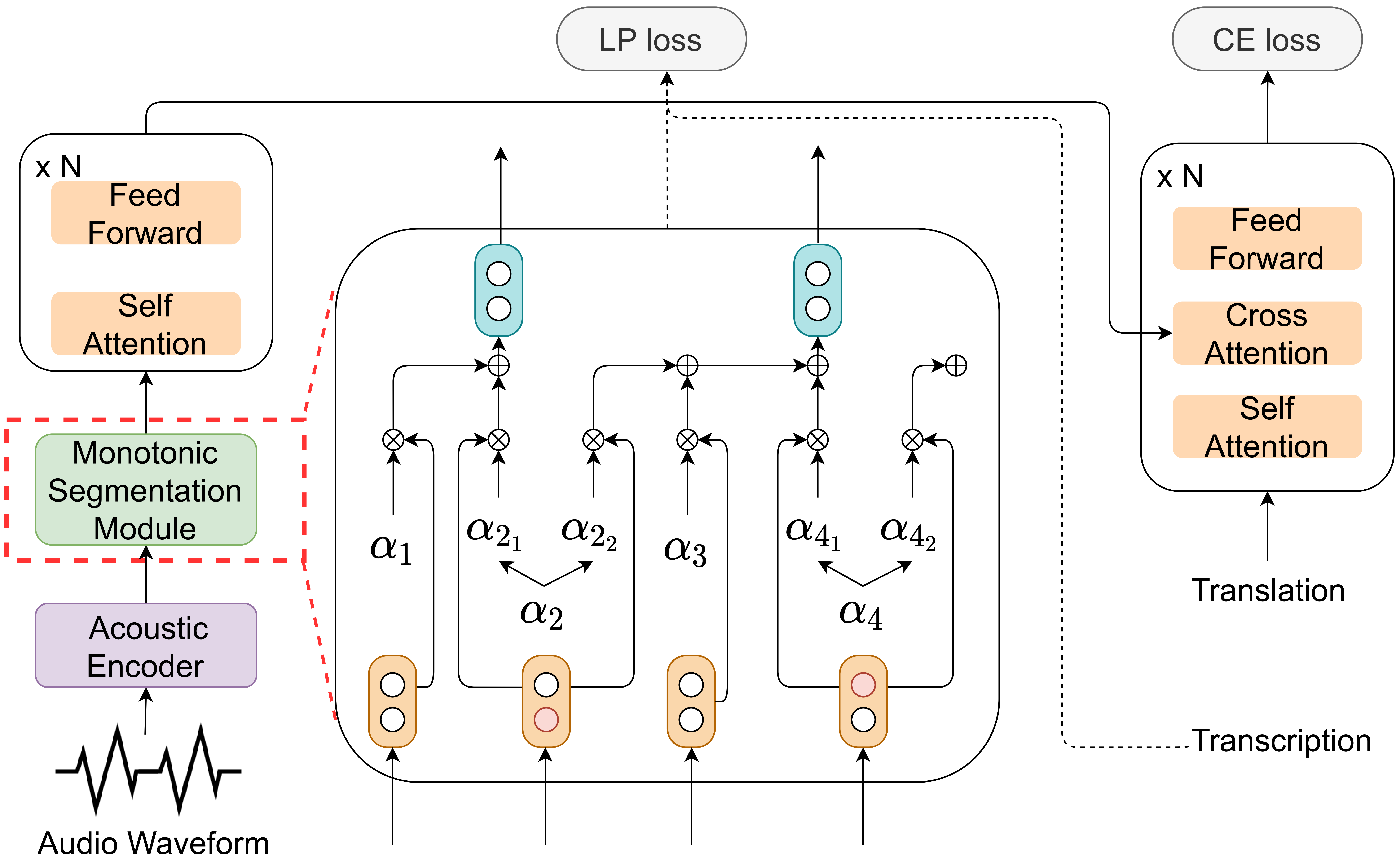
MoSST: Key Insight

Motivation: How to find *proper moments* to generate partial sentence translation given a streaming speech input?

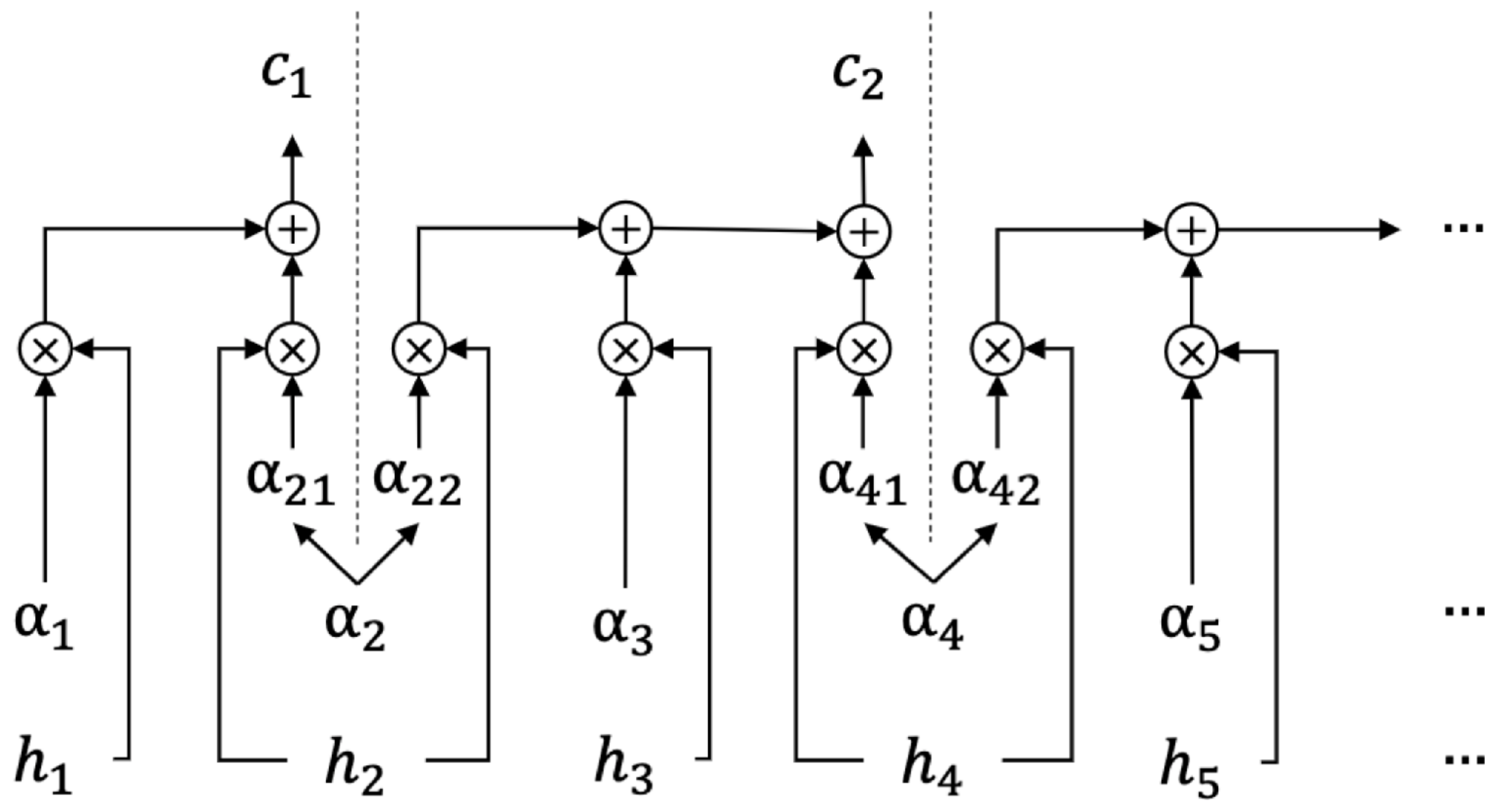
Solution: Introduce a *monotonic segmentation module*.



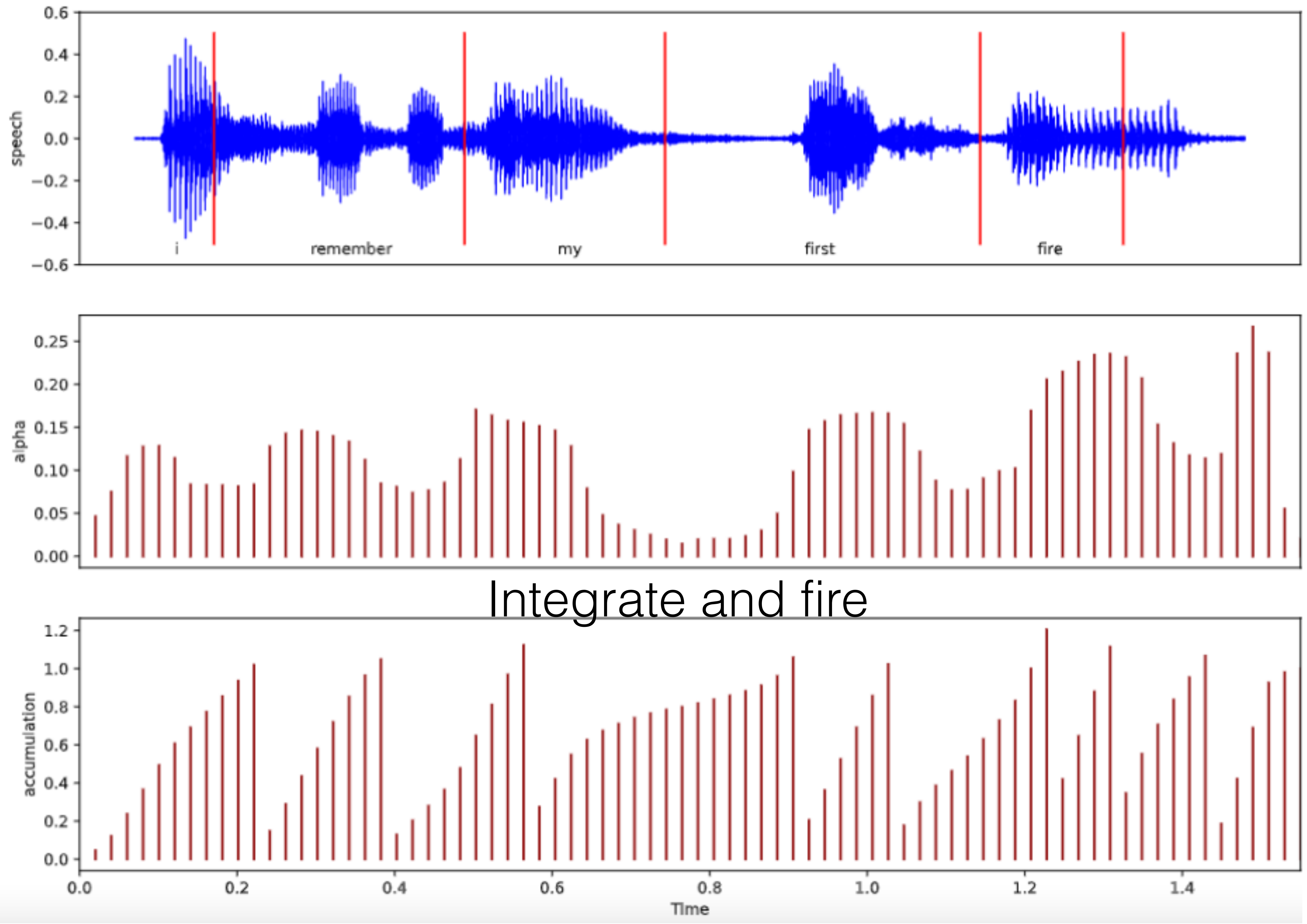
MoSST Overview



Monotonic Segmentation while Listening



(source: Dong et al., 2020[2])

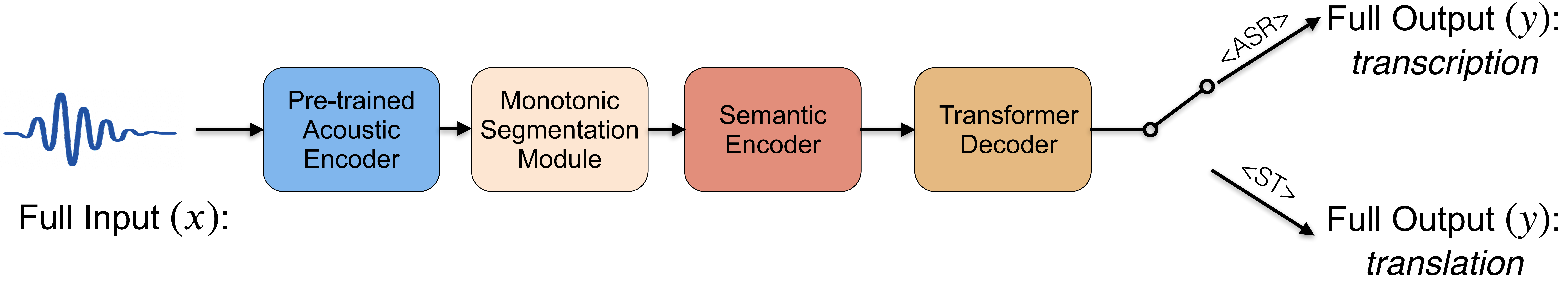


[2] Dong, Linhao, and Bo Xu. "CIF: Continuous integrate-and-fire for end-to-end speech recognition."

MoSST: Training Strategies

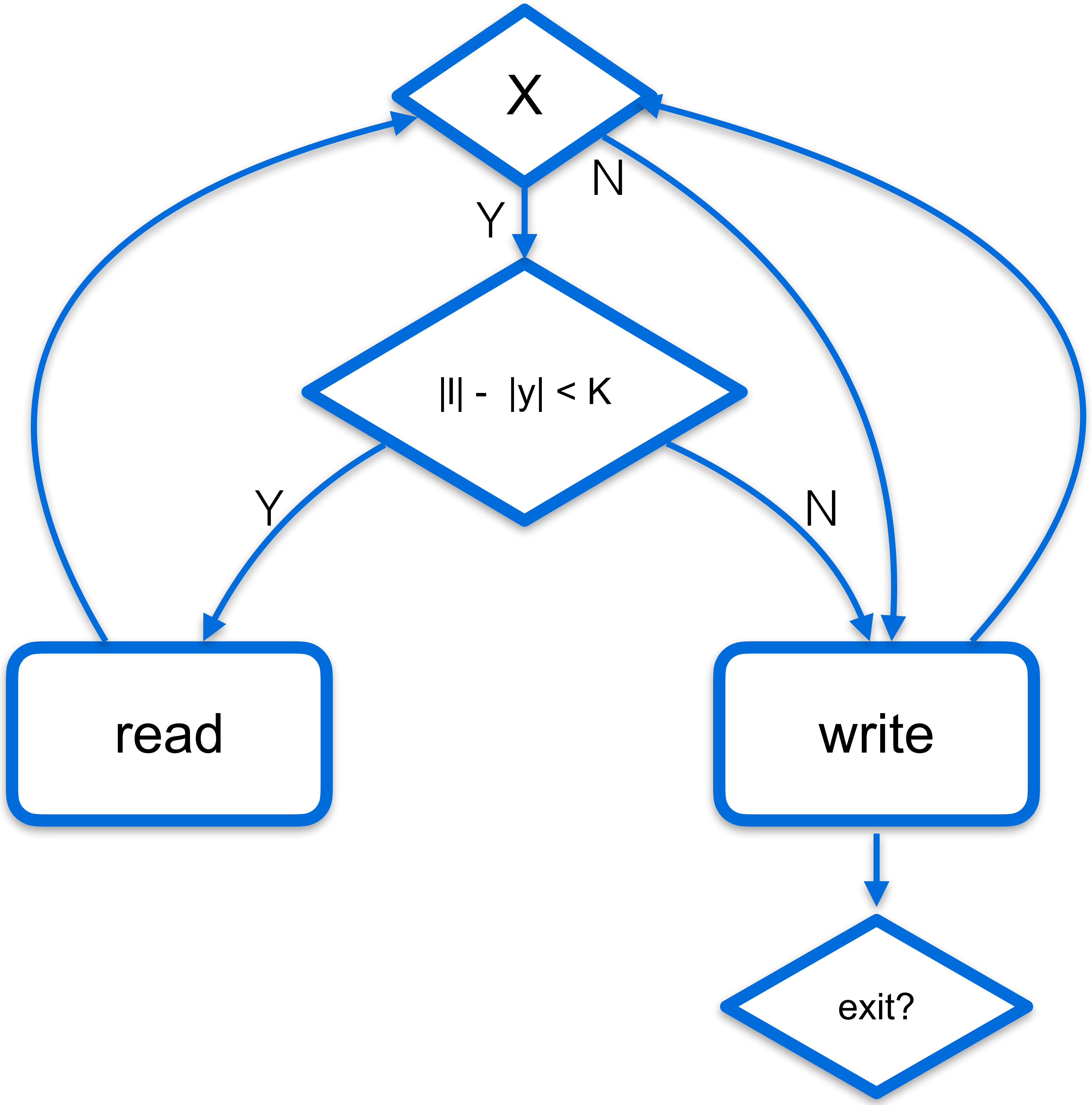
- Full-sentence training without Wait-K is ok!

- To alleviate the data scarcity problem:
  - Pre-trained Acoustic Model
  - Multi-task Training



MoSST Adaptively decide when to Generate Translation

- Adaptive Decision vs Pre-fixed Decision



Experimental Setups

1. Datasets

MuST-C, En?De/Fr

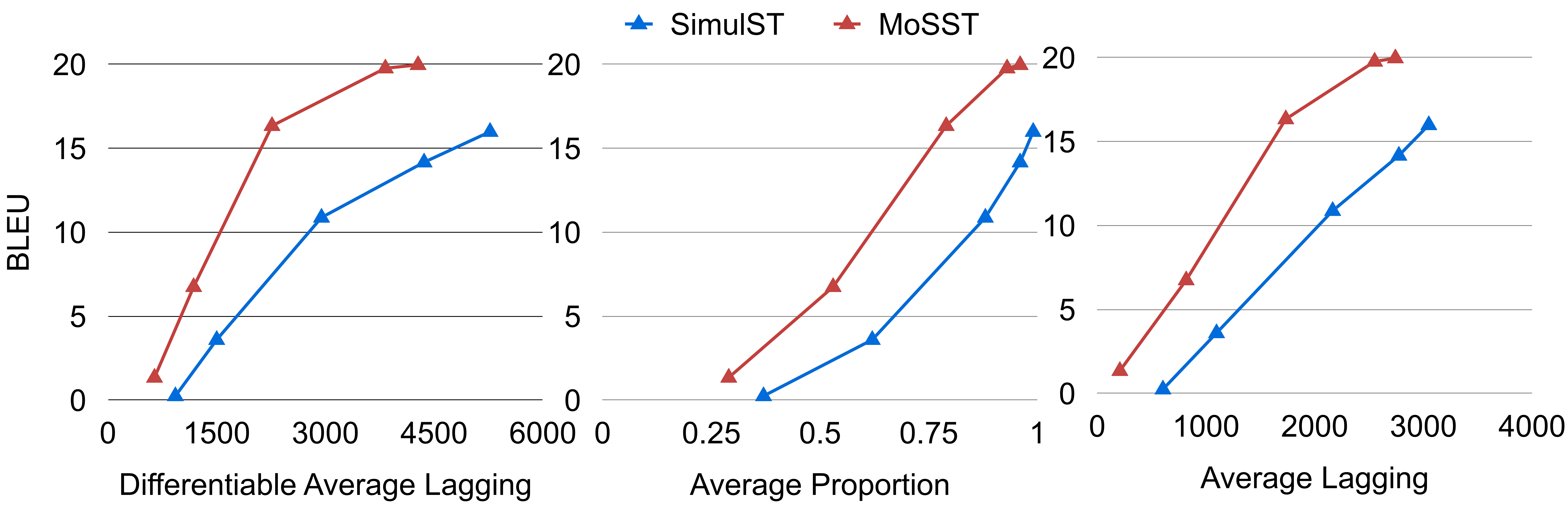
2. Metrics

Accuracy	• BLEU
Latency	• Differentiable Average Lagging • Average Proportion • Average Lagging

3. Model

Module	Backbone
Acoustic Encoder	Wav2vec 2
MSM	**FC**
Semantic Encoder	Transformer Encoder
Decoder	Transformer Decoder

MoSST works much better

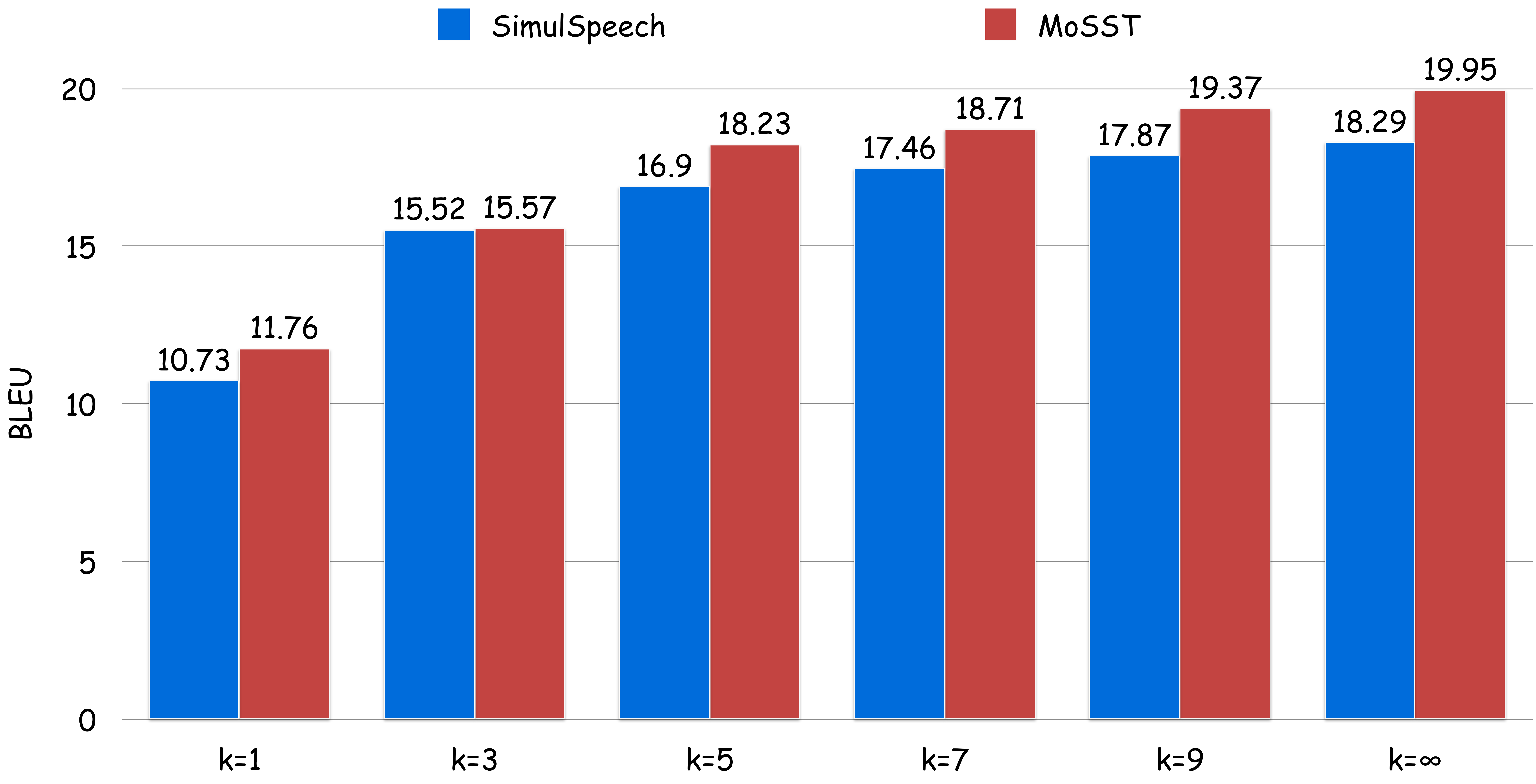


MoSST achieves ***best translation accuracy*** with ***the same lagging***.

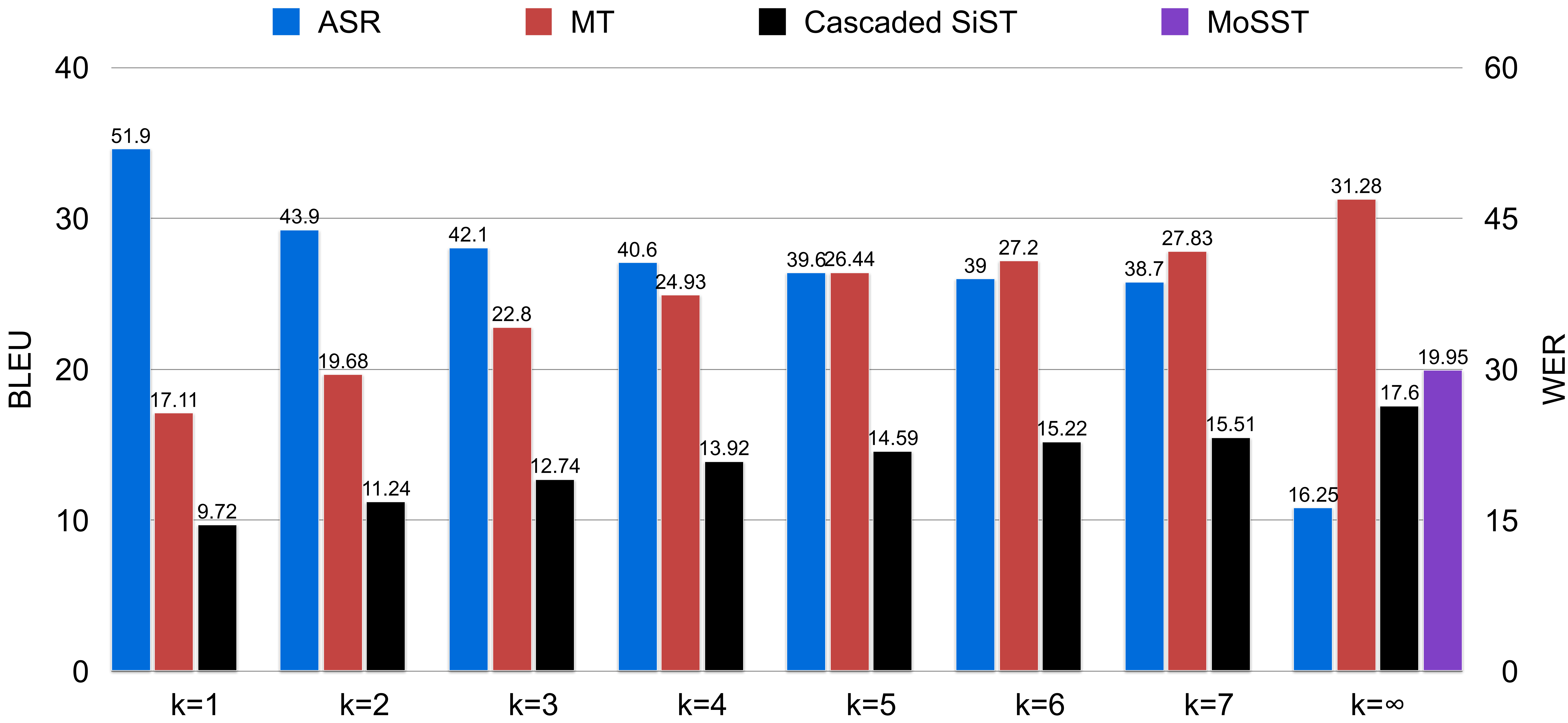
MoSST Is Better Than SimulST (Ma et al., 2020b)

MoSST achieves ***best translation accuracy*** with ***the same lagging***.

MoSST Is Better Than SimulSpeech (Ren et al., 2020)



MoSST Is Superior to Cascaded SiST



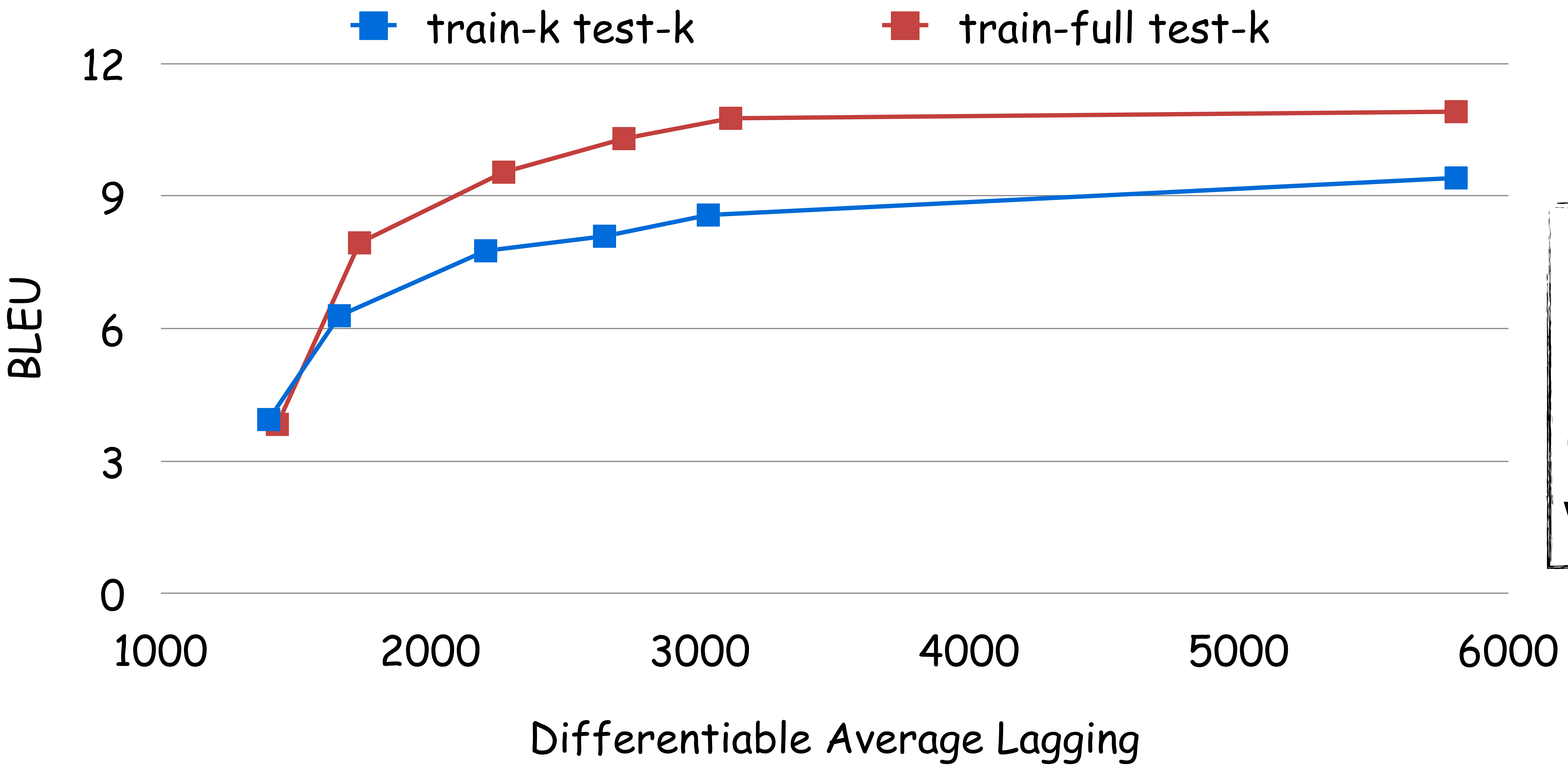
MoSST also works for Offline ST

Model	EN->DE	EN->FR
Transformer ST Fairseq (Wang et al., 2020)	22.7	32.9
Transformer ST ESPnet (Inaguma et al., 2020)	22.9	32.8
Transformer ST NeurST (Zhao et al., 2021)	22.8	33.3
AFS ST (Zhang et al., 2020)	22.4	31.6
STAST (Liu et al., 2020)	23.1	-
Dual-Decoder Transformer (BL) (Le et al., 2020)	23.6	33.5
Wav2Vec2 + Transformer (Han et al., 2021)	22.3	34.3
W-Transf (Ye et al., 2021)	23.6	34.6
RealTrans (Zhang et al., 2021)	22.99	-
MoSST	**24.9**	**35.3**

(under the constrained setting)

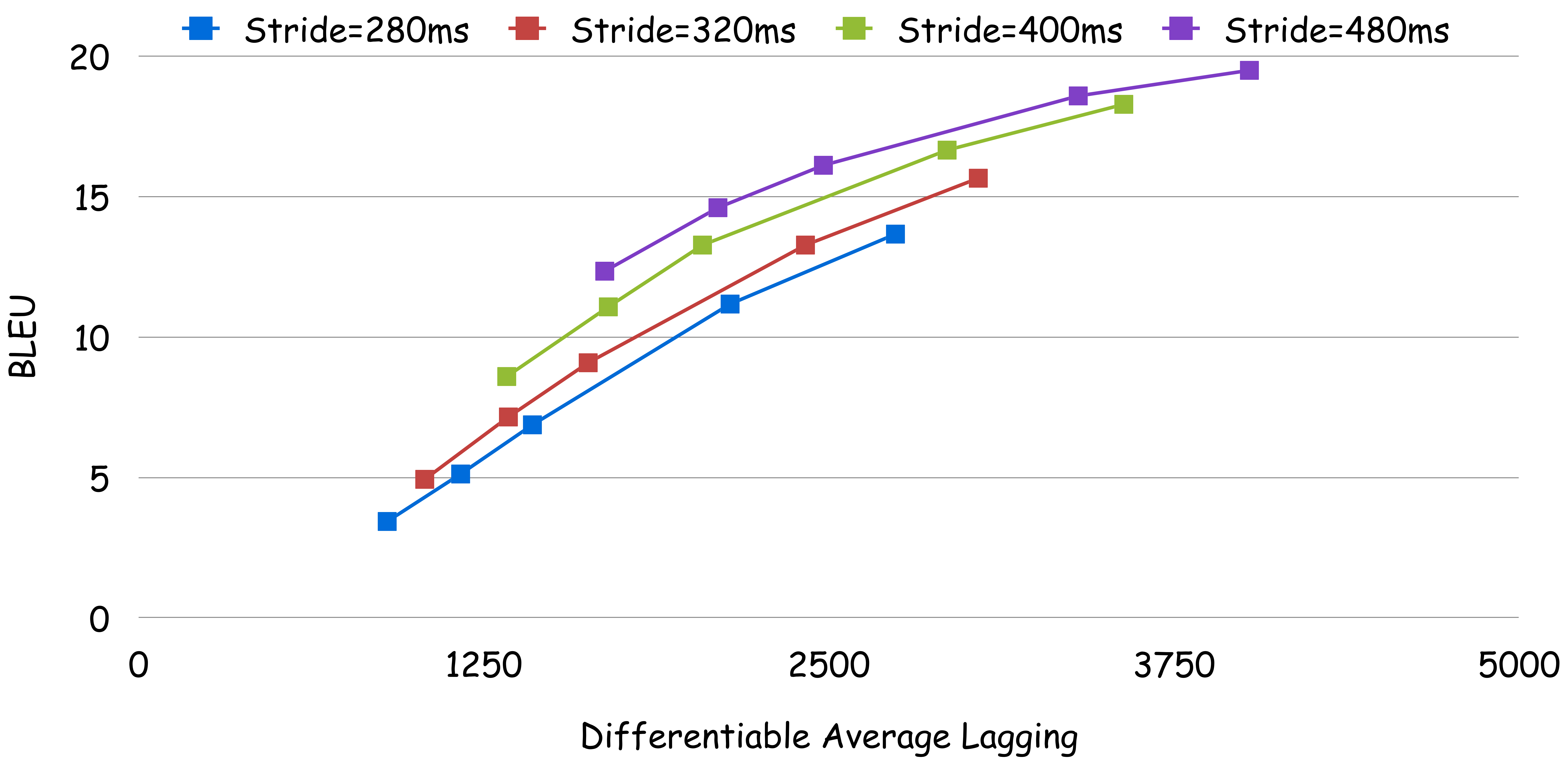
Simultaneous Training (prefix-to-prefix) Is Not Necessary for SiST

- ConvTransformer with offline ASR pre-training

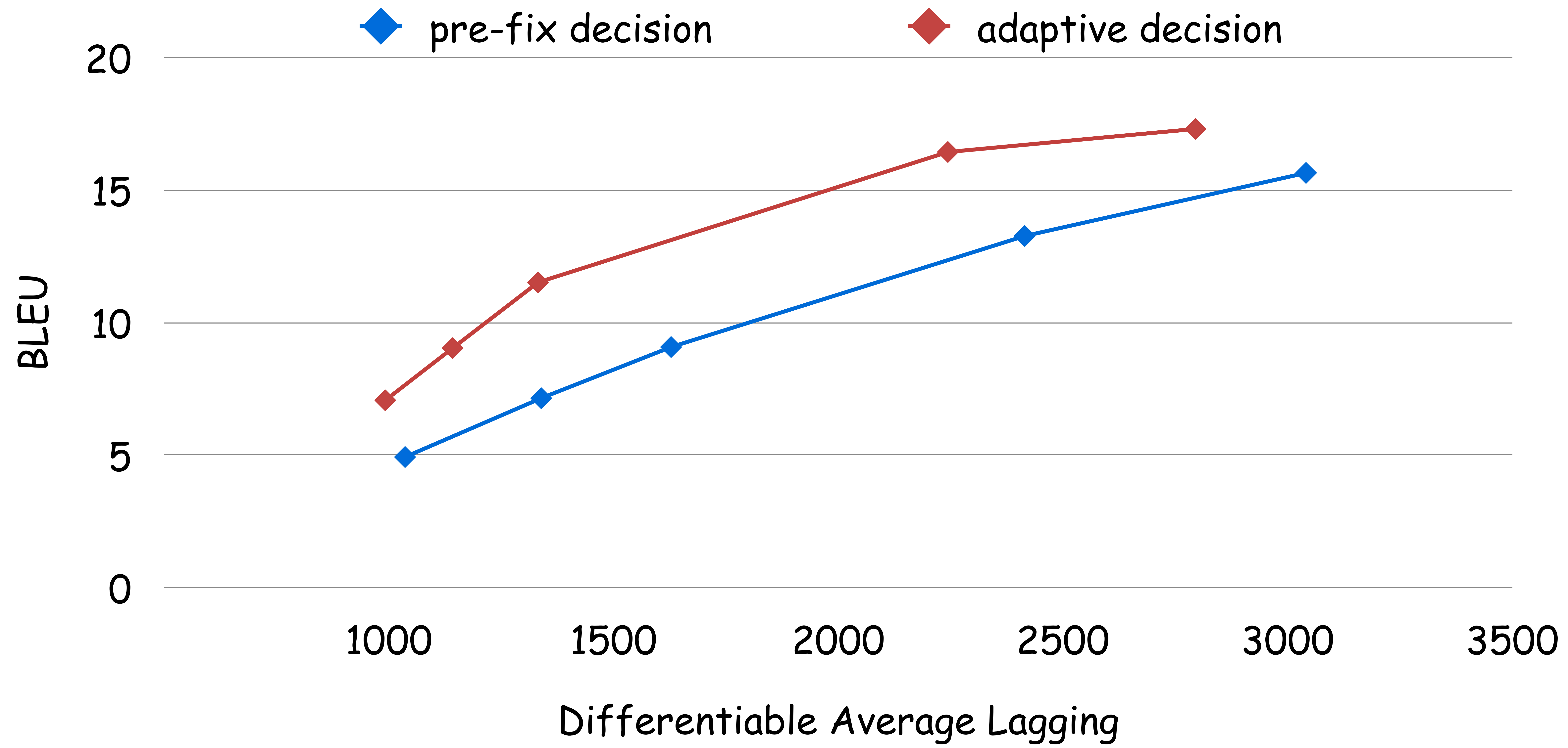


Offine training
is better than
online training
with Wait-K.

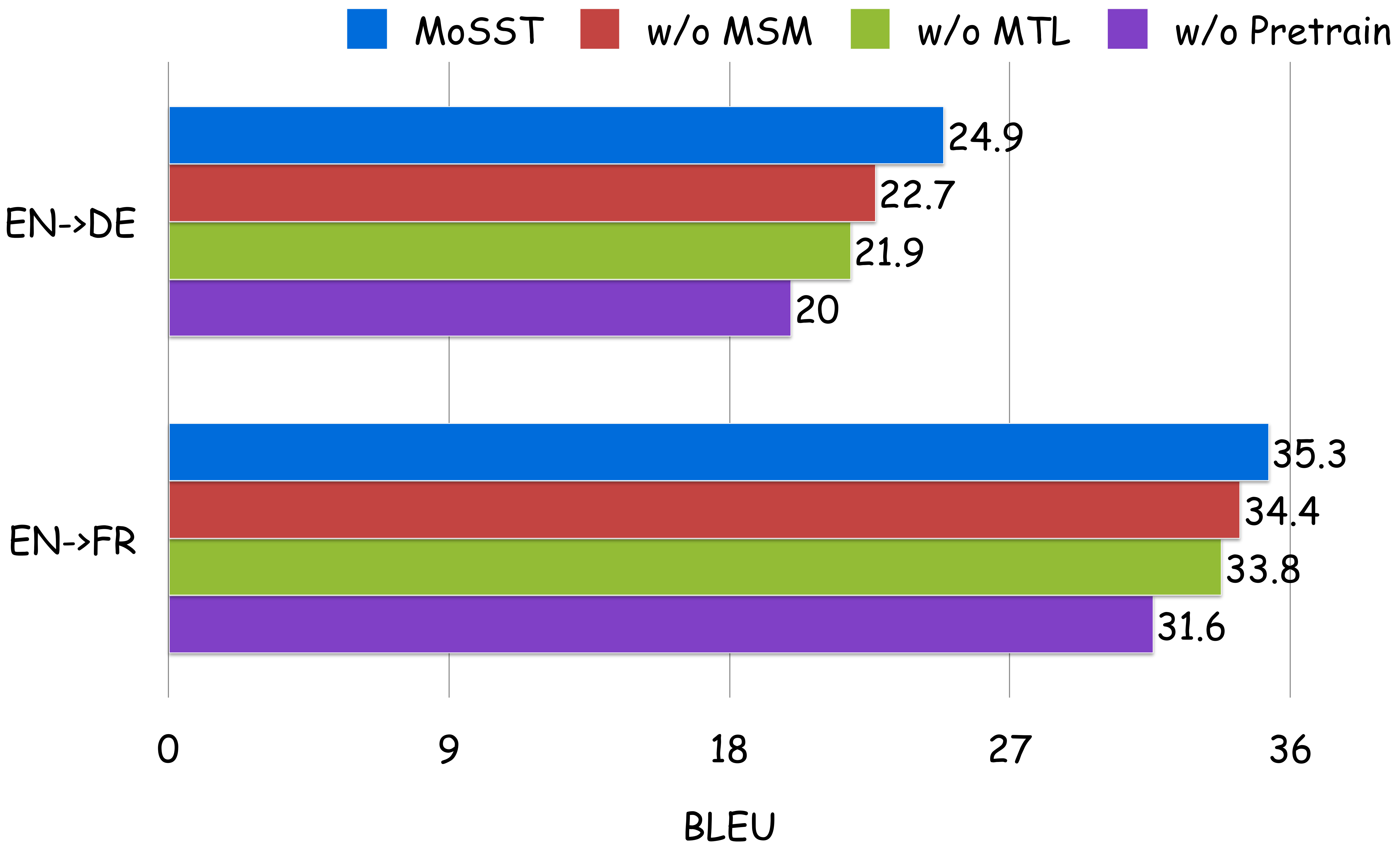
Fixed Strides: Bigger Stride, Higher Latency



Adaptive Decision Is Better Than Pre-fix Decison



Monotonic Segmentation (MSM) is Important!



Case Study

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
En (Source)	If	you	have	somethi ng	to	give	,	give	it	now	.			
De (Target)	Wenn	Sie	etwas	zu	geben	haben	,	geben	Sie	es	jetzt	.		
ASR	If	you	have	somethi ng	to	give	and	give	it	now	.			
Cascades				Wenn	Sie	etwas	zu	geben	haben	und	es	jetzt	geben	.
MoSST				Wenn	Sie	etwas	geben	,	geben	Sie	es	jetzt	.	

Pause in source speech matters!

Takeaways

- End-to-end SiST is a more challenging area that requires balancing accuracy and latency.

- To segment audio waveform into acoustic units, MoSST introduces a new monotonic segmentation module, based on which the adaptive decision strategy can dynamically decide when to translate in streaming scenarios.

- MoSST can significantly outperform SOTA baselines both for streaming and non-streaming ST.

Language in 10