End-to-End Simultaneous Speech Translation with Pretraining and Distillation: Huawei Noah’s System for AutoSimTranS 2022

Xingshan Zeng, Pengfei Li, Liangyou Li, Qun Liu

Huawei Noah’s Ark Lab

{zeng.xingshan,lipengfei111,liliangyou,qun.liu}@huawei.com

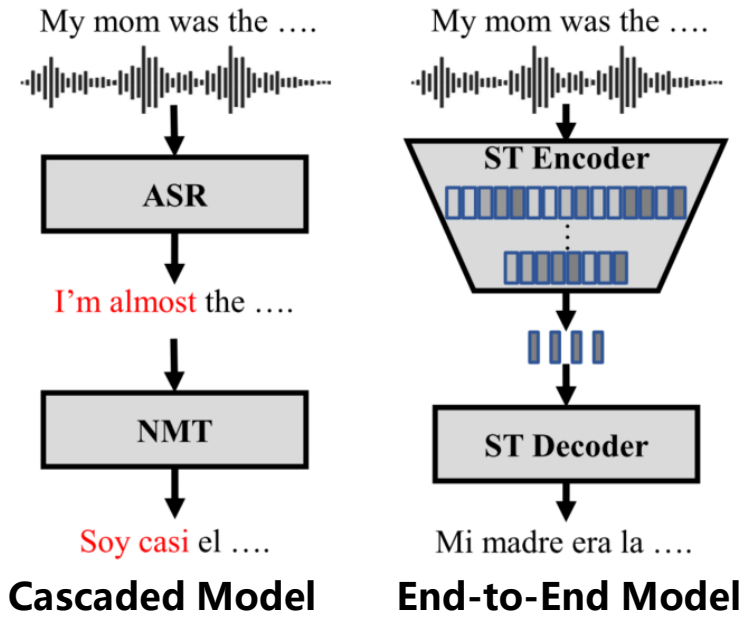
HUAWEI

Introduction

- E2E Speech Translation Definition:
  - Directly translating speech in one language into text in another language in real-time.

- Advantages (vs. cascaded model):
  1. Small model
  2. Low latency
  3. Avoid error propagation

- Problems:
  - **Data scarcity**
  - Modality gap
  - Controllability



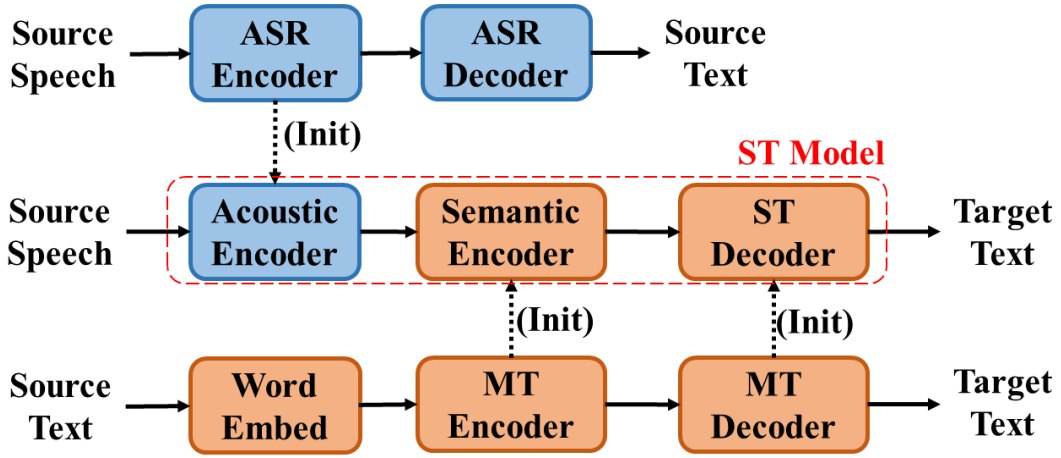
Cascaded Model **End-to-End Model**

Data Name	Data Type	#Hours	#Sents
BSTC	ST	70	38K
Wenet Speech	ASR	10K	14M
WMT17	NMT	--	9M

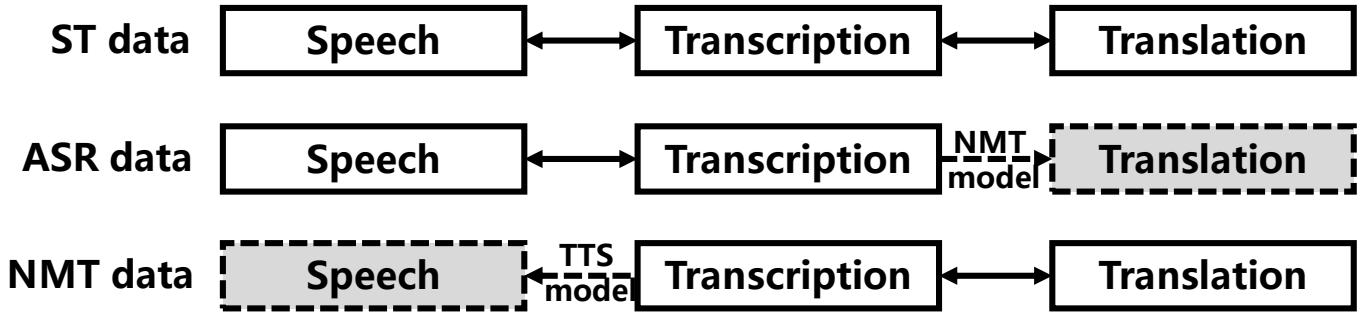
Zh-En Data

Introduction

- Pretraining
  - Using pretrained ASR and NMT model to initialize the modules of ST model

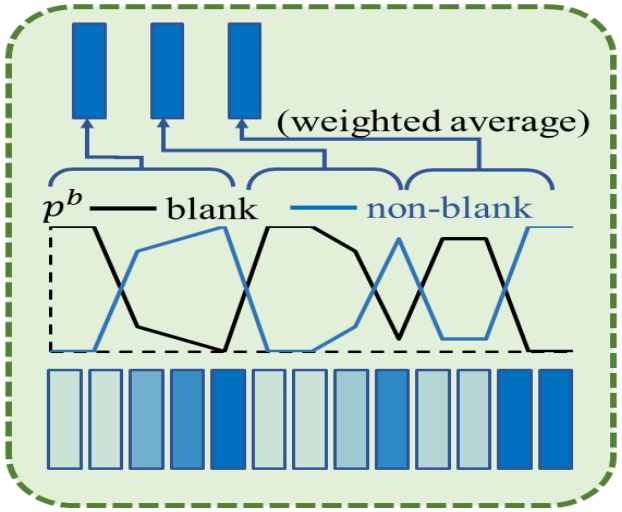


- Distillation
  - Constructing Pseudo data
    - ASR data + NMT model
    - NMT data + TTS model

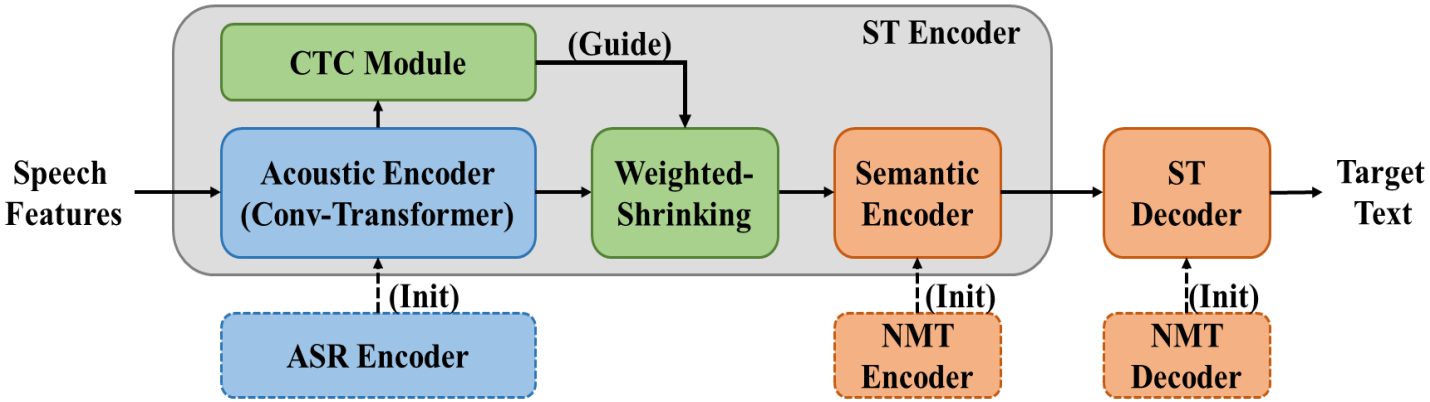


Introduction

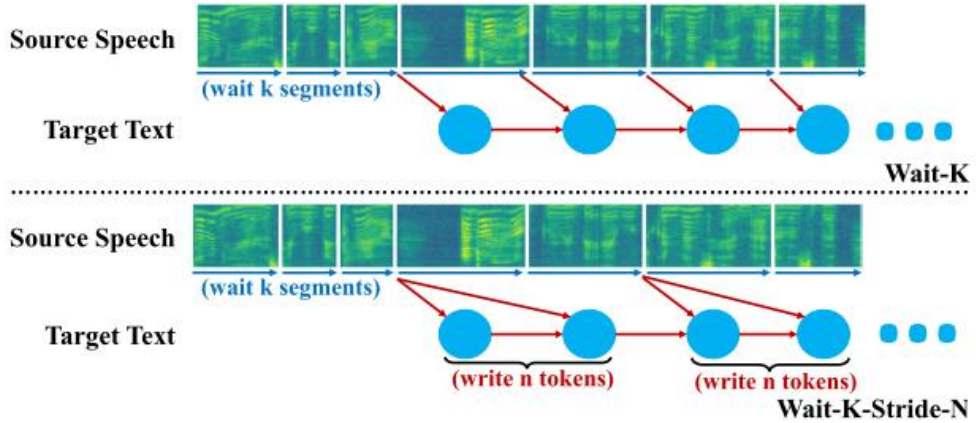
- Model: RealTranS (Zeng et al. 2021)
  - 8x downsampling + weighted shrinking
  - Blank-limited CTC module
  - Wait-K-Stride-N policy
    - Wait-K: first wait K segments (detected by CTC)
    - Stride-N: each time write N tokens (then read N segments)



Weighted Shrinking



Model Overview



Wait-K vs Wait-K-Stride-N

HUAWEI

Training Procedure

1. Pretraining
   - ASR pretraining: ConvTT (Huang et al. 2020)
   - NMT pretraining: CeMAT (Li et al. 2022)
2. Training on Pseudo Data
   - ASR data + NMT model -> ST data
   - Multi-path wait-k training
   - Punctuation removal
3. Finetuning with In-Domain Data
   - Multi-domain finetuning
   - Token-level knowledge distillation
   - Punctuation removal

Dataset	SRC Speech	SRC Text	TGT Text	#Hours	#Sents
CWMT21	×	✓	✓	–	9M
Internal	✓	✓	Pseudo	10K	11M
WenetSpeech	✓	✓	Pseudo	10K	14M
BSTC	✓	✓	✓	70	38K

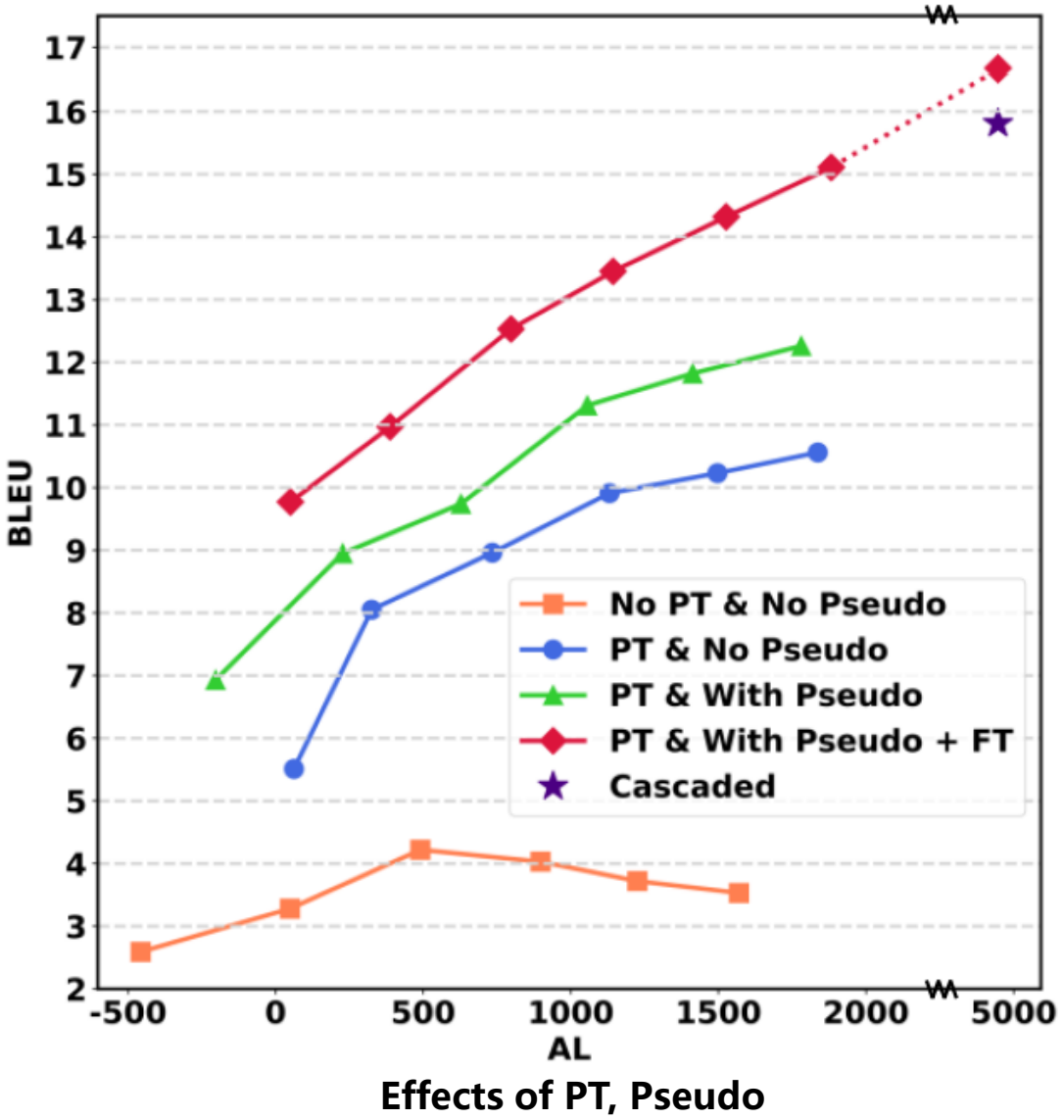
Used Data

Punctuation Removal:
Removing punctuation in transcription.

This is to ensure consistency among different datasets, as many of ASR datasets do not include punctuation.

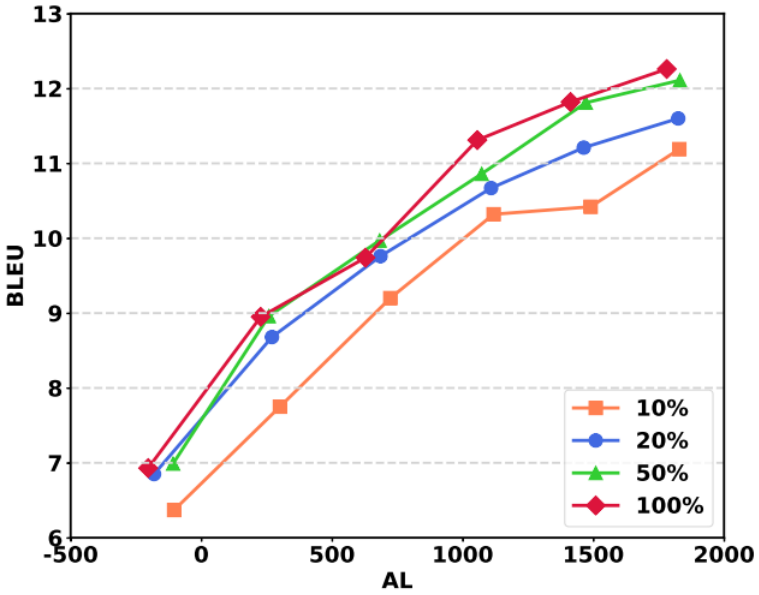
Experiments

a. No PT & No Pseudo: directly train the model on the in-domain data.
b. PT & No Pseudo: use pretraining and directly train the model on the in-domain data.
c. PT & With Pseudo: use pretraining and train the model on the pseudo data.
d. PT & With Pseudo + FT: further finetune model c on the in-domain data.
e. Cascaded: results by cascading our pretrained ASR and NMT model.



More Experiments

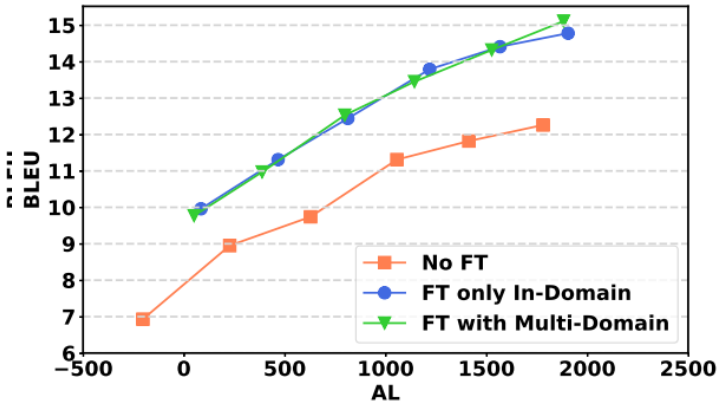
- Some findings:
  - Effects of pseudo data are limited by the ability of NMT model.
  - Multi-domain finetuning is beneficial for domain generalization.
  - Punctuation removal helps on acoustic learning and improves performance by keeping consistent between pretraining and finetuning.



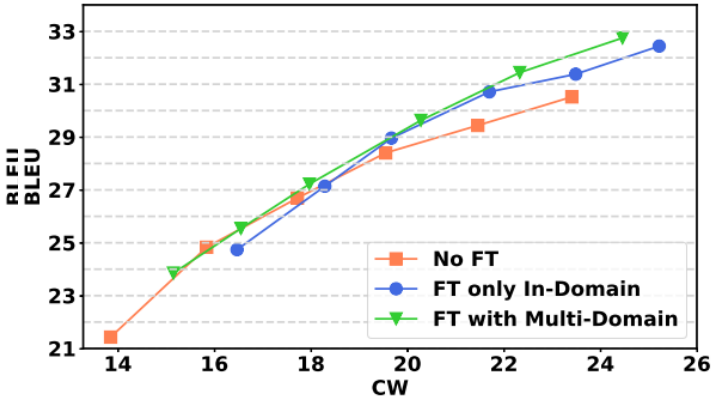
Effects of Pseudo Data Amount

Model	CTC Loss (↓)		BLEU (↑)	
	With Punct	RM Punct	With Punct	RM Punct
No PT & No Pseudo	4.88	4.60	4.12	4.03
FT based on PT + Pseudo	2.10	1.73	11.81	12.41

Effects of Punct Removal



Dev Set



Test Set

Effects of Multi-Domain Finetuning

Thank you!