NAIST Simultaneous Speech Translation System for IWSLT 2023

Ryo Fukuda[1], Yuta Nishikawa[1], Yasumasa Kano[1], Yuka Ko[1], Tomoya Yanagita[1],
Kosuke Doi[1], Mana Makinae[1], Sakriani Sakti[21], Katsuhito Sudoh[1], Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology, Japan [2]Japan Advanced Institute of Science and Technology, Japan

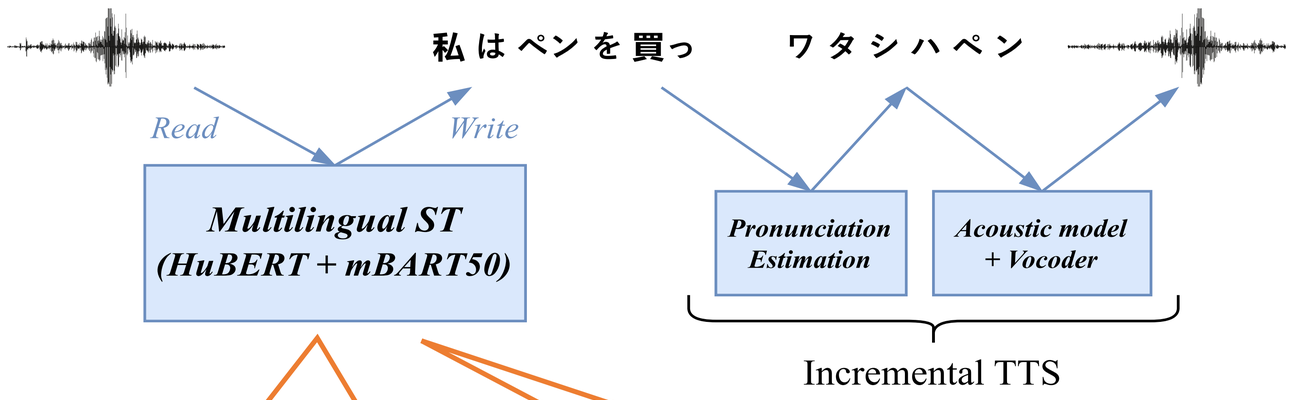
Our Systems

■ SimulS2T for En-{De,Ja,Zh}

- Model: Multilingual ST with pretrained models
  - Encoder: HuBERT with Inter-connection
  - Decoder: mBART50 decoder layers
- Training: Two-stage fine-tuning
  1. Multilingual ST corpora
  2. Prefix pairs extracted using Prefix Alignment (PA)
- Policy: Local Agreement [Liu+2020]

■ SimulS2S for En-Ja

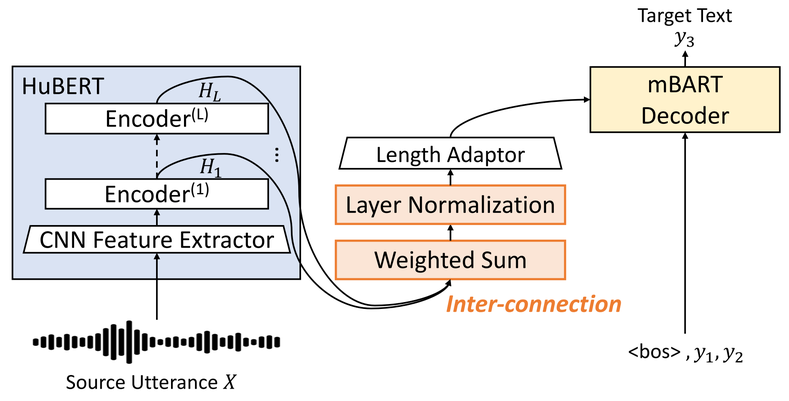
- Pronunciation Estimation
  - subword to Japanese pronunciation (katakana)
  - LSTM with wait-2
- Acoustic model + Vocoder
  - Tacotron2 and neural vocoder



Inter-connection [Nishikawa+INTERSPEECH2023]

Effective connection between pre-trained Encoder and Decoder

➢ Aggregate hidden states from intermediate layers of HuBERT
→ Input it to the mBART Decoder



Prefix Alignment [Kano+IWSLT2022]

Data augmentation for prefix-to-prefix translation

➢ Extract prefix pairs using offline translation
→ fine-tune an offline ST model for SimulST

Source Prefix	Prefix Translation (*gloss*)	Offline translation
I	私は。(*I.*)	私はペンを買った
I bought	私は買った。(*I bought*)	
I bought a	私は一つ買った (*I bought one*)	
I bought a pen	私はペンを買った (*I bought a pen*)	

Prefix pairs

[(“I”, “私は”),
(“I bought pens.”, “私はペンを買った。”)]

SimulS2T

■ Setup: Local Agreement with $n = 2$ (LA-2), $chunk\ size = 200{\sim}100$ ms

➢ **PA-fine-tuned model outperformed an offline baseline**

- The best En-Ja results came from filtering 73% of prefix pairs
  ⇒ Reducing unbalanced pairs in distant languages is important
- Filtering did not work well for En-De and En-Zh

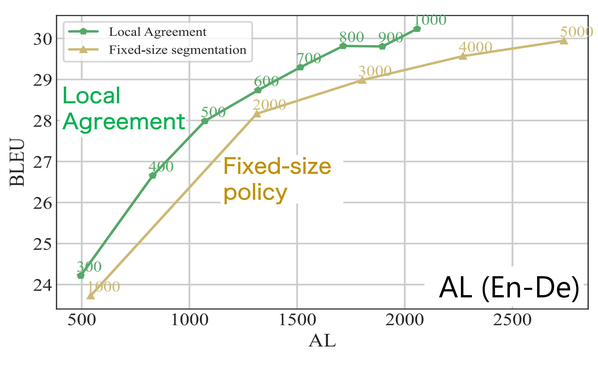
➢ **Inter-connection worked for En-De and En-Ja**
⇒ Shared aggregation weights were helpful across multiple languages

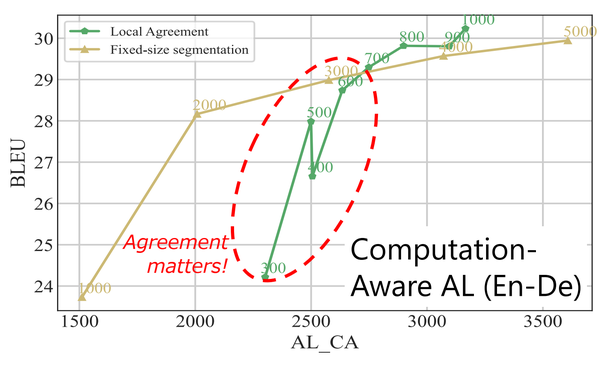
Results with and w/o Inter-connection.

Model	En-De	En-Ja	En-Zh	Ave.
HuBERT + mBART	30.47	15.71	**25.01**	23.73
w/ Inter-connection	**30.89**	**15.89**	24.75	**23.84**

➢ **Fixed-size policy can be a good choice in practice.**

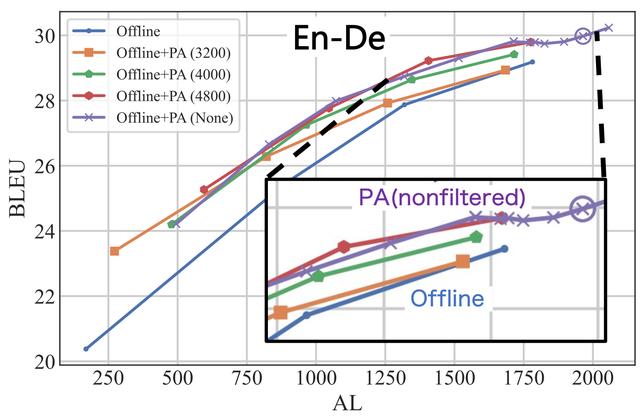
- Local Agreement was faster than fixed-size policy in non-computation-aware AL
- Local Agreement was slower than fixed-size policy in computation-aware AL

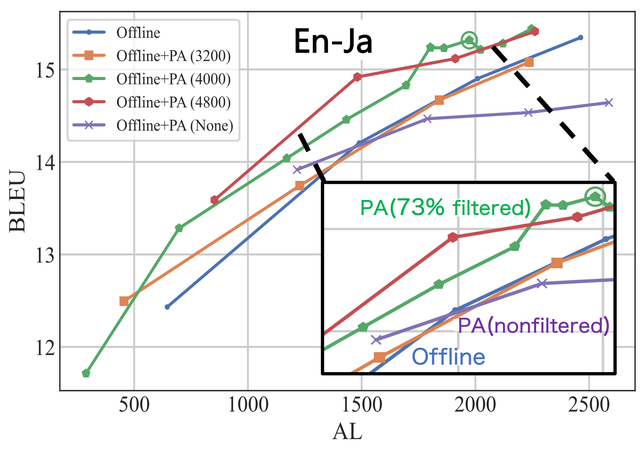




Submitted S2T systems on MuST-C v2 tst-COMMON.

Lang pair	chunk size	BLEU	AL
En-De	950 ms	29.98	1964
En-Ja	840 ms	15.32	1974
En-Zh	700 ms	22.11	1471





SimulS2S

■ Setup: ASR_BLEU with Whisper-large

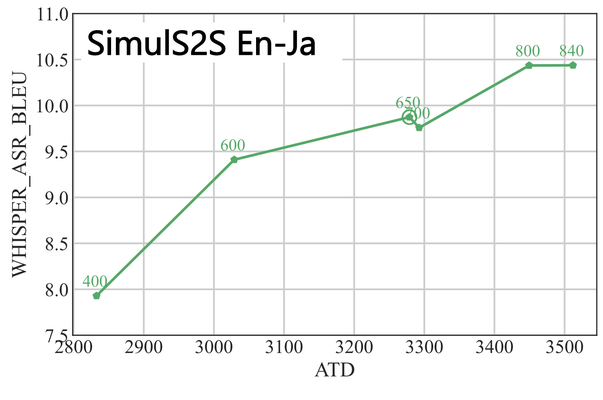
➢ **SimulS2S resulted in much worse compared to S2T En-Ja**

- BLEU 15.32 → ASR_BLEU 9.87
- S2S had a character error rate of 28.3%
- Japanese TTS has difficulty controlling intonation

⇒ Significant room for improvement in the TTS

Submitted S2S system on MuST-C v2 tst-COMMON.

ASR_BLEU	StartOffset	EndOffset	ATD
9.87	2495	4135	3279



Summary

■ SimulS2T results showed

- effectiveness of Prefix Alignment and Inter-connection.
- superiority of fixed policy in computation-aware latency.

■ SimulS2S results showed

- promising performance by simple cascade of SimulS2T and incremental TTS.