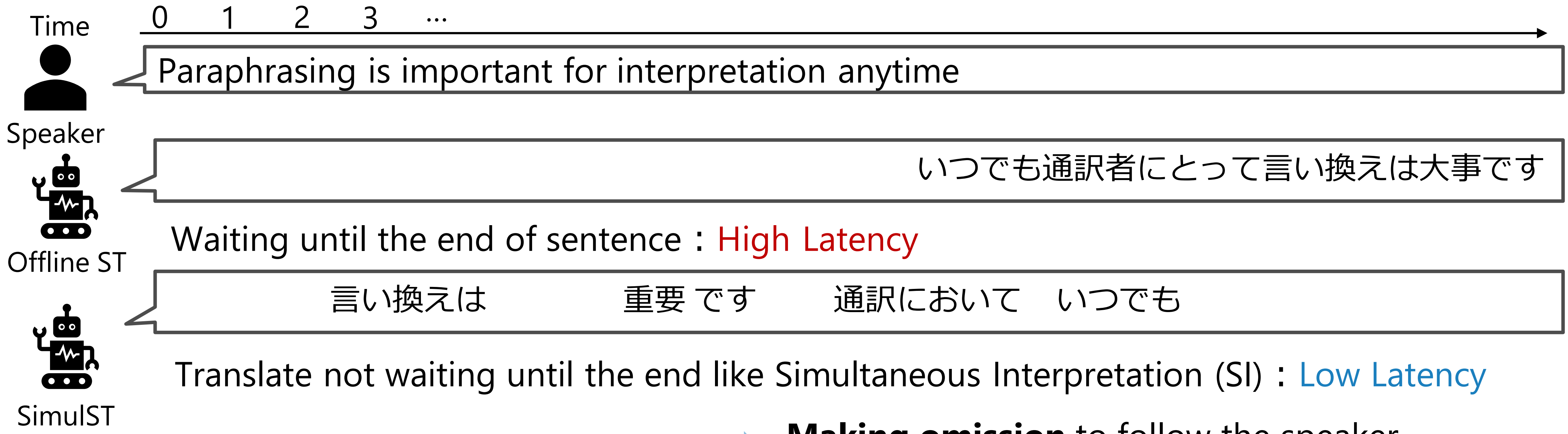
Tagged End-to-End Simultaneous Speech Translation Training using Simultaneous Interpretation Data

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano,
Katsuhito Sudoh, Satoshi Nakamura
ko.yuka.kp2@is.naist.jp
Nara Institute of Science and Technology (NAIST), Japan

IWSLT 2023 7/13-7/14

Background: Simultaneous Speech Translation (SimulST)



Paraphrasing is important for interpretation anytime

いつでも通訳者にとって言い換えは大事です

Waiting until the end of sentence：High Latency

言い換えは 重要 です 通訳において いつでも

Translate not waiting until the end like Simultaneous Interpretation (SI)：Low Latency

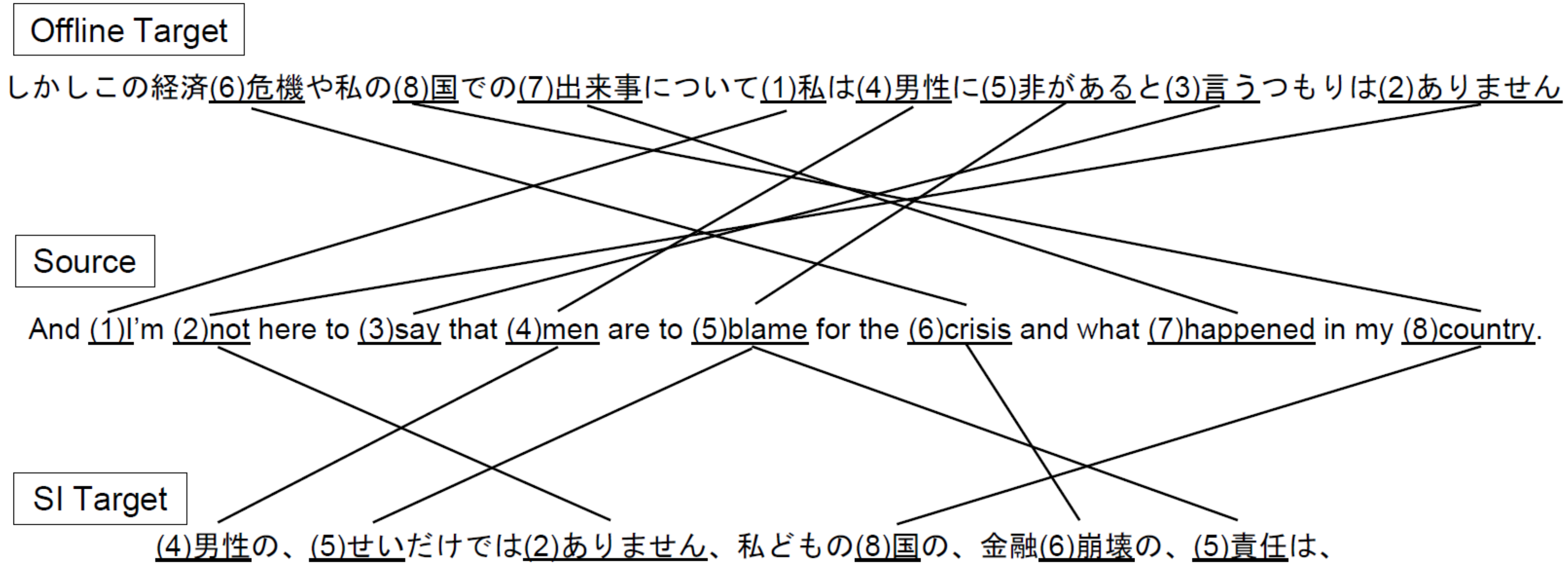
SI-like output using SI data → **Making omission** to follow the speaker
Translating in **monotonic order**

- Problem
  - Fine-tuning (FT) with SI data causes overfitting in small SI data
- This work
  - Using both offline data and SI data
  - Controlling output style with style tags

Background: Offline and SI output

- Offline
  - English words have correspondences of Japanese
  - Keeping naturalness with long-distance reordering
- SI
  - Some words are dropped or omitted
  - Translating in monotonic order

→ Follow the speaker's speech
Generate the words earlier



Offline Target

しかしこの経済(6)危機や私の(8)国での(7)出来事について(1)私は(4)男性に(5)非があると(3)言うつもりは(2)ありません

Source

And (1)I'm (2)not here to (3)say that (4)men are to (5)blame for the (6)crisis and what (7)happened in my (8)country.

SI Target

(4)男性の、(5)せいだけでは(2)ありません、私どもの(8)国の、金融(6)崩壊の、(5)責任は、

Related work

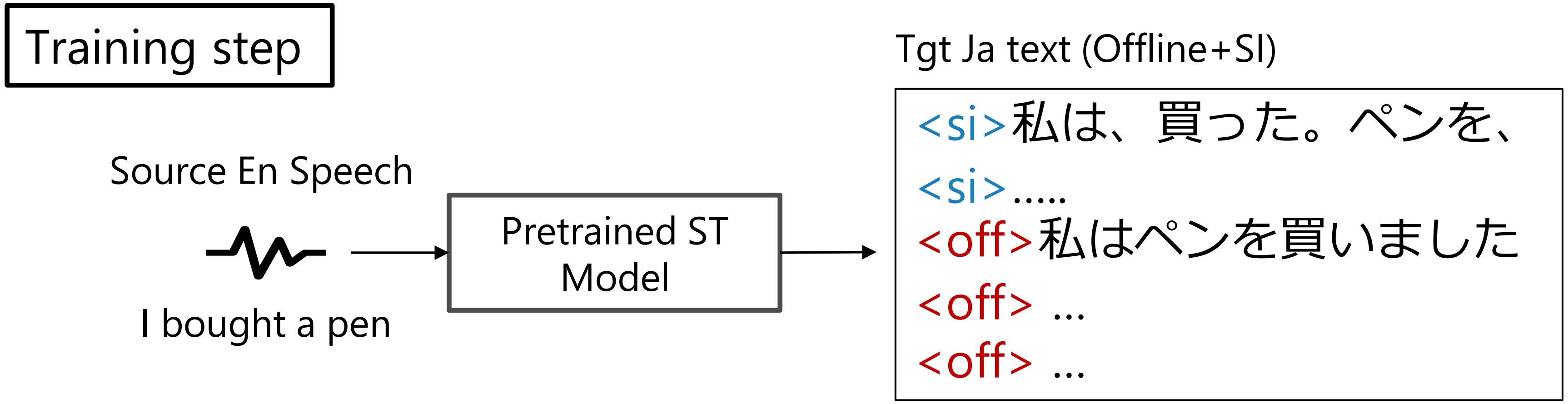
- SI corpora (in English-Japanese)
  - SI data in English-Japanese Small amount of SI data
    - Not sentence-to-sentence aligned data [Toyama+2004, Shimizu+2013, Doi+2021]
    - Sentence-to-sentence aligned data [Zhao+2023]

- Domain adaptation using tags For small data training
  - Mixed fine-tuning with out-domain and in-domain [Chu+2017] avoids overfitting
  - Tag-based NMT [Sennrich+2016]
  - Zero-shot multilingual NMT [Johnson+2017]
  - Tagged back-translation [Caswell+2019]

Proposed Method: Mixed FT with style tags

- Training
  - Putting style tags at the beginning of target texts
- Inference
  - Decoding in forced decoding with prefix style tags

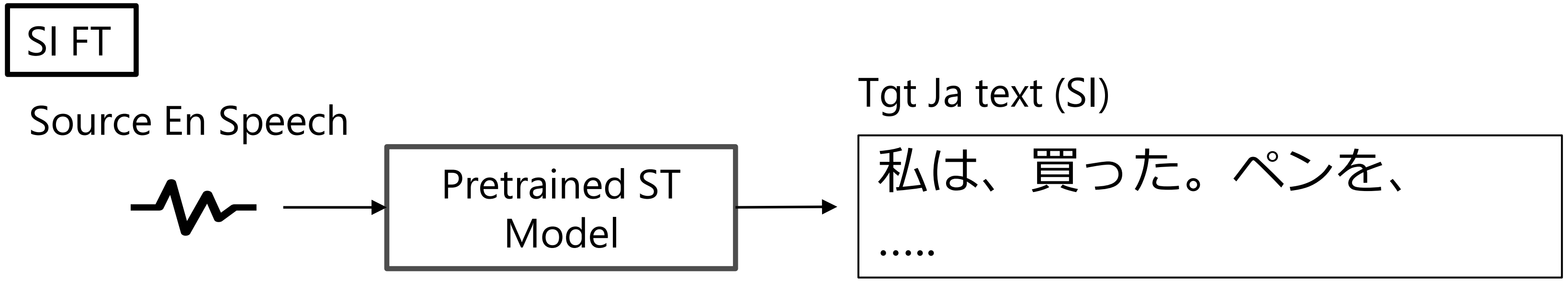
Motivation:
- Mitigating SI data scarcity problem avoiding overfitting
- Using large offline and small SI data effectively



<si>私は、買った。ペンを、
<si>.....
<off>私はペンを買いました
<off> ...
<off> ...

Experiment setting

- **Baseline Model**
  - **Offline FT**
  - **SI FT**
  - **Mixed FT**



SI FT

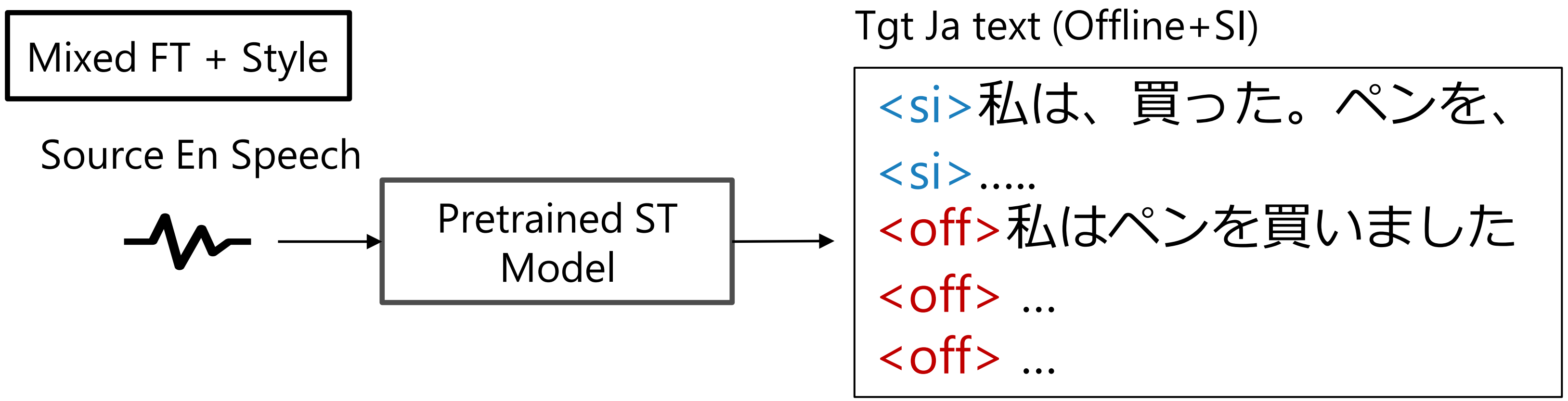
Source En Speech

Pretrained ST Model

Tgt Ja text (SI)

私は、買った。ペンを、
.....

- **Proposed Model**
  - **Mixed FT + Style**: Fine-tuning with both offline and SI data with style tags
  - **Mixed FT + Style + Up**: Up-sampling in SI data



Mixed FT + Style

Source En Speech

Pretrained ST Model

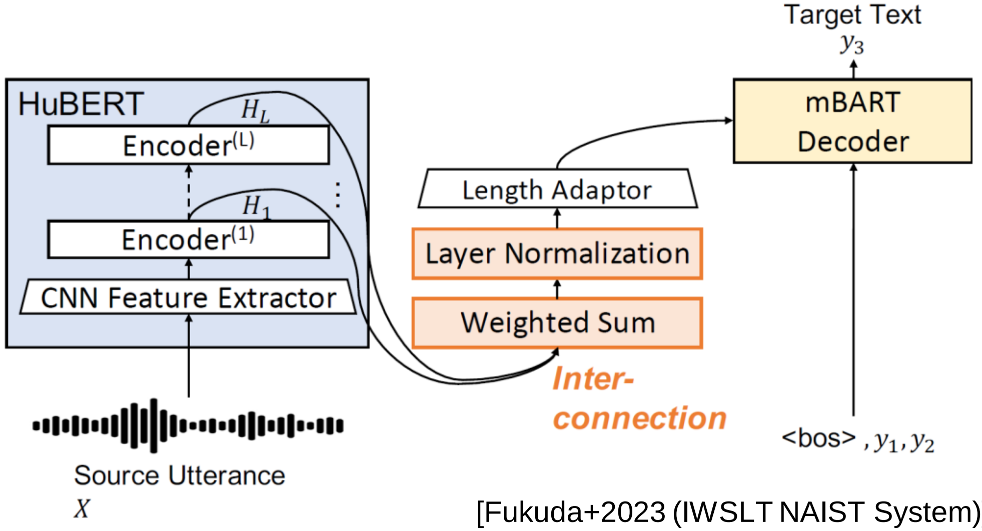
Tgt Ja text (Offline+SI)

<si>私は、買った。ペンを、
<si>.....
<off>私はペンを買いました
<off> ...
<off> ...

Experiment setting

	Offline	SI
Train	328639	65008
Dev	1369	165
Test	2841	511

- Data
  - Offline: MuST-C En-Ja [Di gangi+2019]
  - SI: NAIST-SIC-Aligned INTRA En-Ja [Zhao+2023] for ST*
- Pretrained offline ST model
  - HuBERT+mBART model [Fukuda+2023]
- Simultaneous decoding
  - Local Agreement [Liu+2020]
  - Speech segment size**: {200, 400, 600, 800, 1000}ms
  - Style tag in inference step
    - SI Test: output from <si> tag
    - Offline test: output from<off> tag
- Evaluation metrics
  - SimulEval
    - BLEURT in ATD [Kano+2023]
    - BLEU in ATD



[Fukuda+2023 (IWSLT NAIST System)]

<u>BLEURT</u>
- The **sentence semantic similarity** between hypothesis and reference

<u>ATD (Average Token Delay)</u>
- Latency metric **focuses on the end timings of partial translations**

* We aligned English text segments with corresponding audio in MuST-C with force aligner gentle

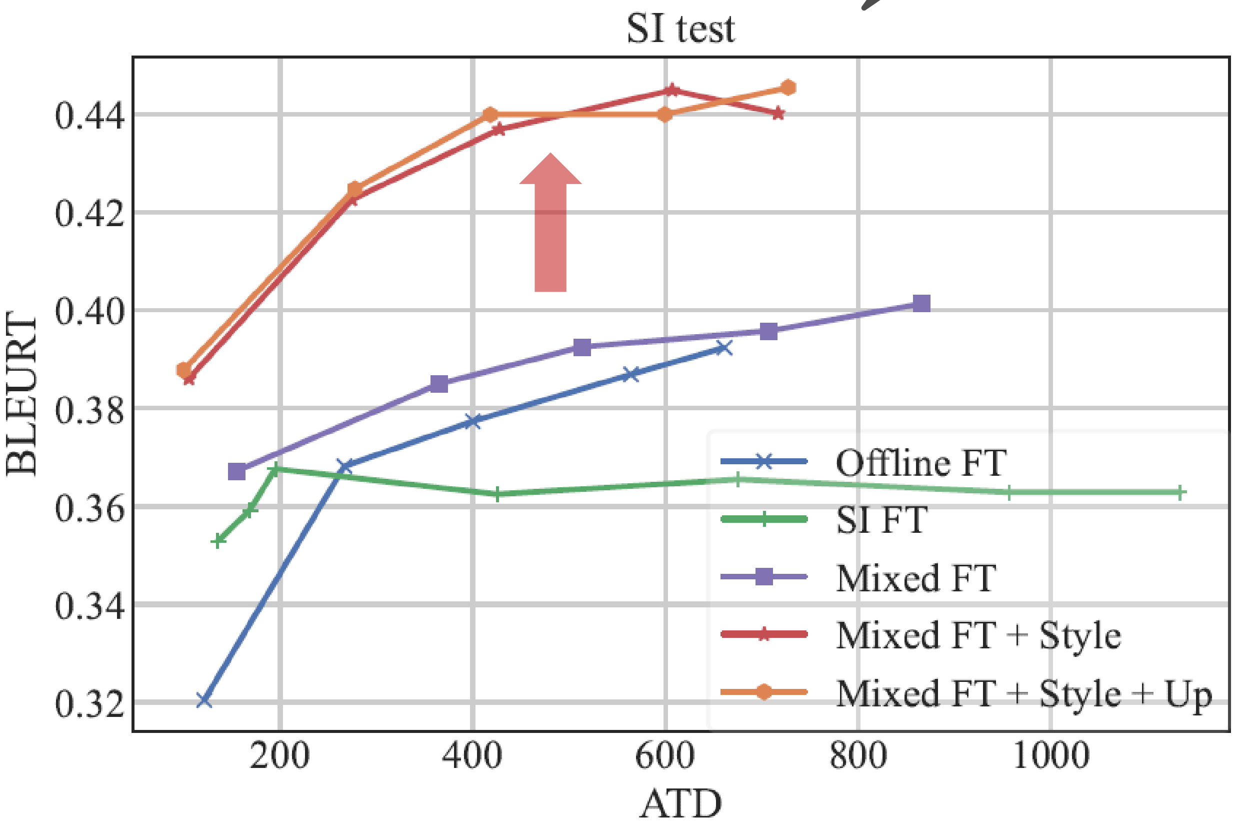
** We also applied 120ms and 160ms for SI FT to see the trend in low latency regime

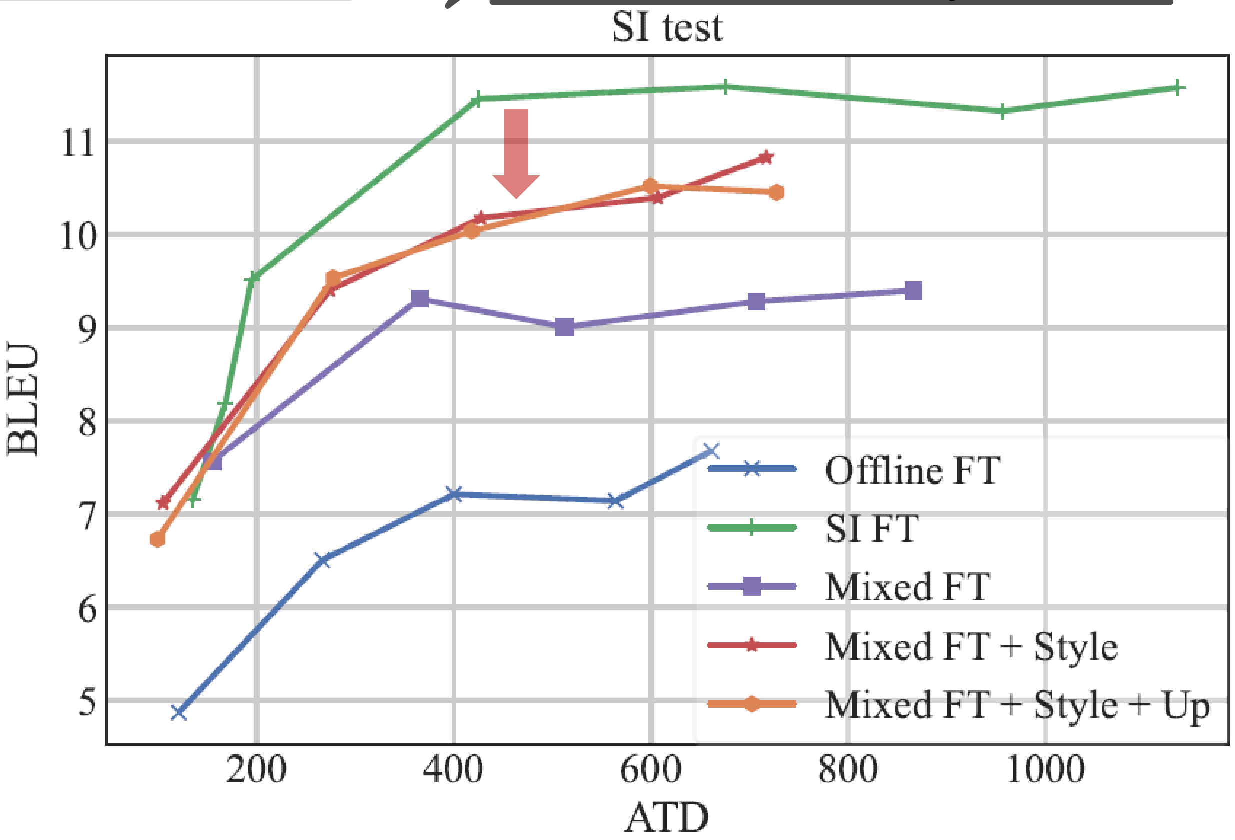
Main results in SI test

- **BLEURT** sentence similarity between hypothesis and reference : **Mixed FT Style** > **SI FT**
- **BLEU** **SI FT** > **Mixed FT Style**

Proposed models were better in sentence similarity

Why the proposed models were lower than baselines in BLEU ? → Next Analysis

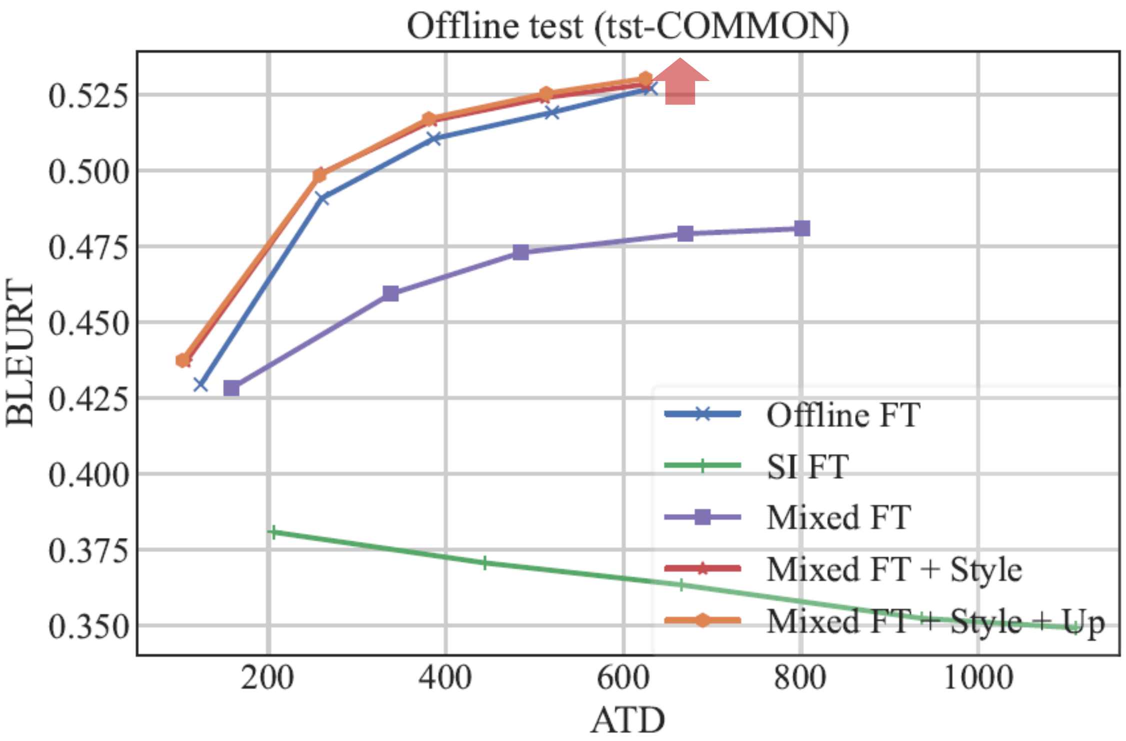


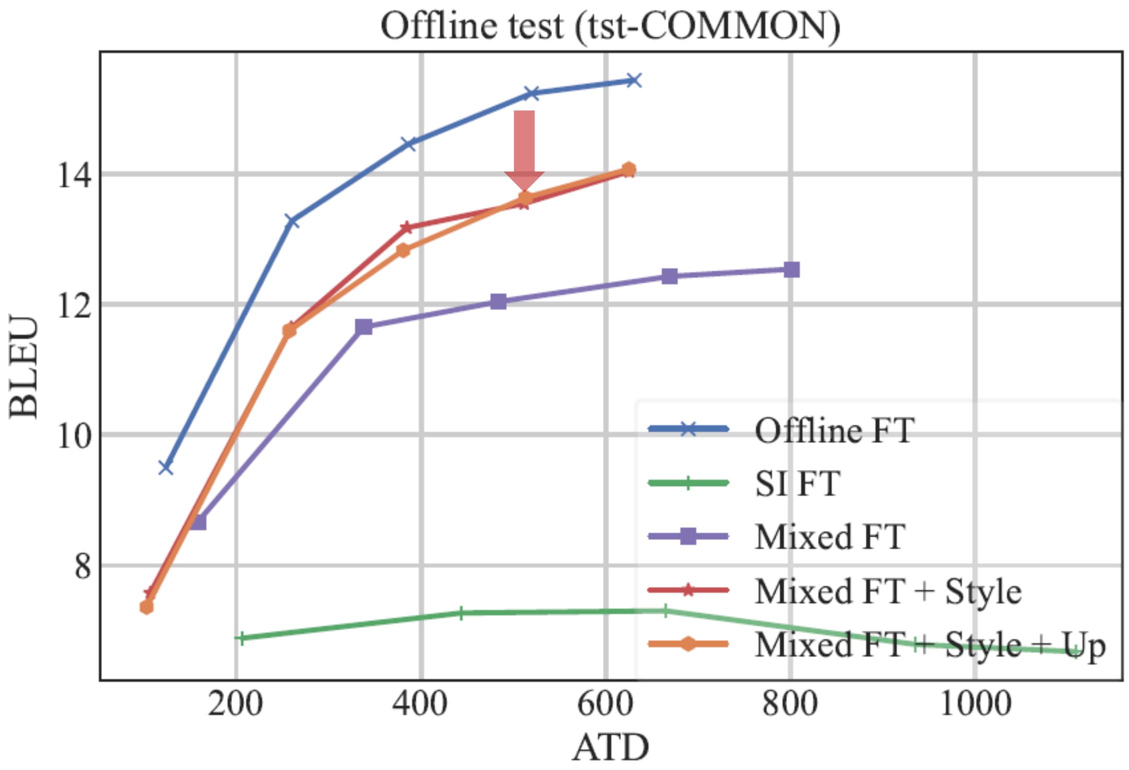


Main results in Offline test (tst-COMMON)

- Offline test
  - The trend was the same as SI test
    - BLEURT: Offline FT < Mixed FT Style
    - BLEU: Offline FT > Mixed FT Style

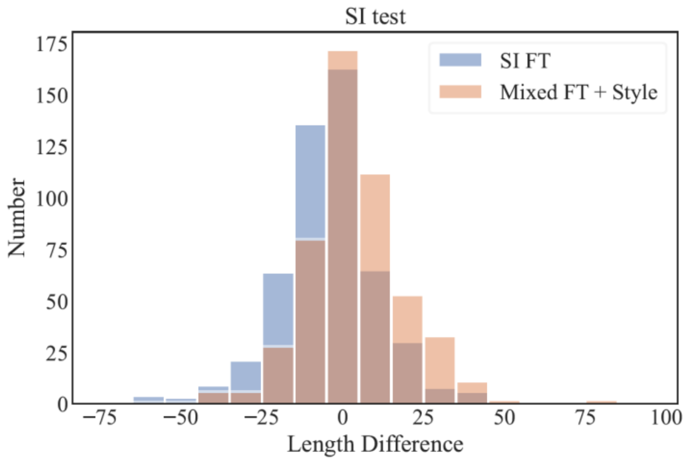
Our model can generate not only SI-like output but also offline-like output





Analysis: output length

- Why proposal (**Mixed FT Style**) < baseline (**SI FT**) in BLEU
- High precision in **SI FT**
  - Small output → tend to be high BLEU
- High recall：**Mixed FT Style**
  - Long output → tend to be low BLEU



SI test	八十年代の素晴らしいグラフィックアーティストでした。TEMPT was one of the foremost graffiti artists in the 80s. 病院も、ノートは言えない。There's no hospital that can say "No." 麻痺してる人達は、これを全員使うことが出来るようになっています。Anybody who's paralyzed now has access to actually draw or communicate using only their eyes.
SI FT (Baseline)	テンプトは、グラフィティアーティストの TEMPT was, graffiti artists' 病院は、a hospital 麻酔した人達は、paralyzed people
Mixed FT + Style (Propose)	テンプとは、グラフィティアーティストの一人です。TEMPT is one of graffiti artists' 病院では「いいえ」は言えません。In a hospital, we cannot say "No." 麻痺した人なら誰でも、絵を描いたり、会話をすることができます Anybody who is paralyzed can draw a picture and have a talk.

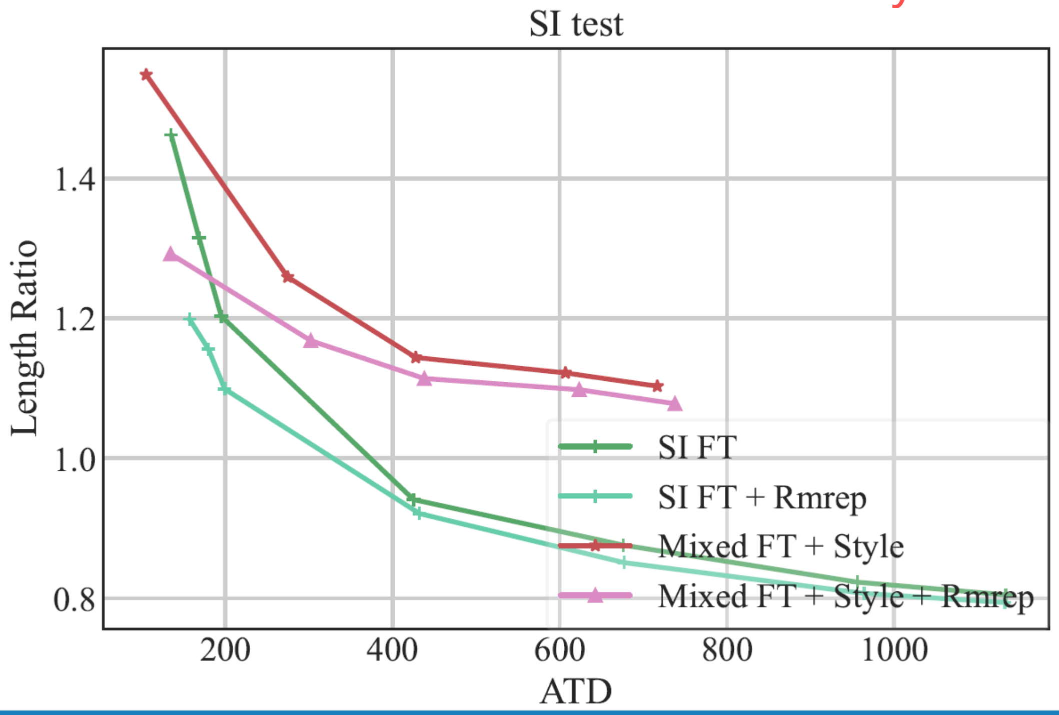
SI FT: Lacking the information included in SI test reference

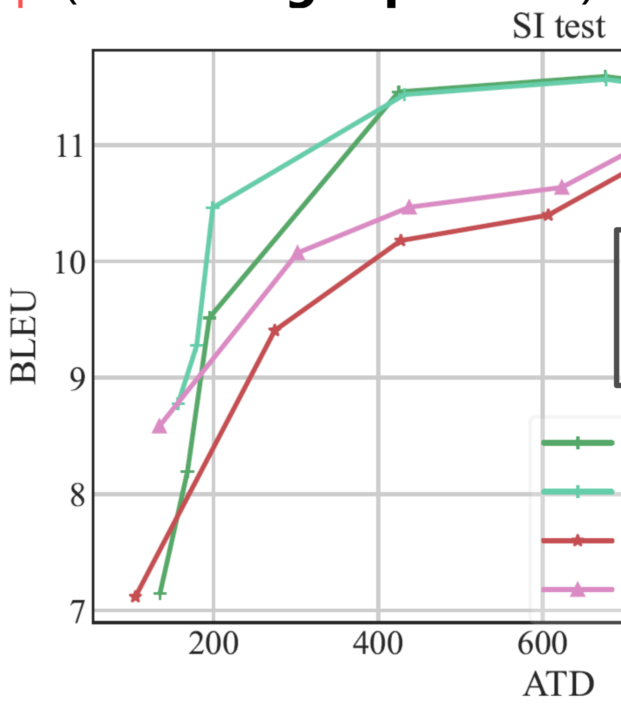
Analysis: repetition by non-speech sound event label

- Why **Mixed FT Style** was generating long output?
  - Repetitions from non-speech sound event label
  - There was repetitions like (Laugher) (Laugher) ... In Japanese → Long output trend → resulted in low BLEU
    - Offline ST tgt text: included
    - SI tgt text: excluded

⇨ Resulted in repetition in proposed models

- Additional evaluation: Mixed FT + Style + Rmrep (**removing repetition**)

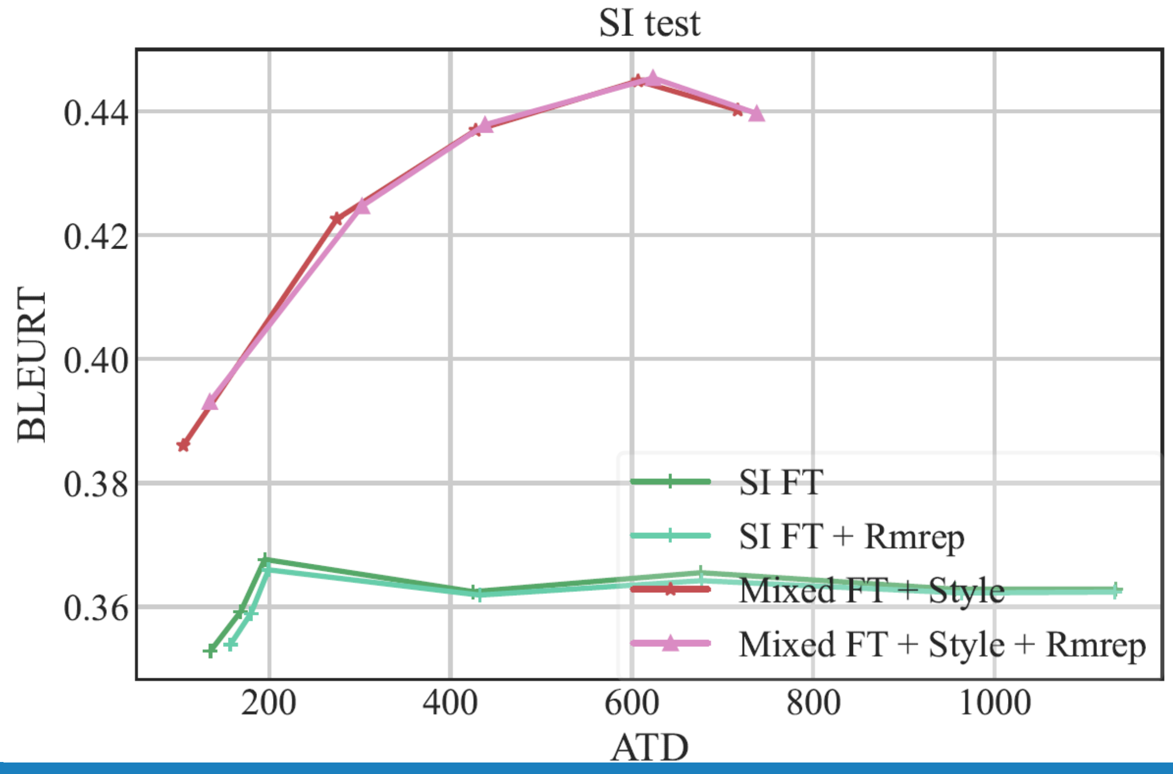


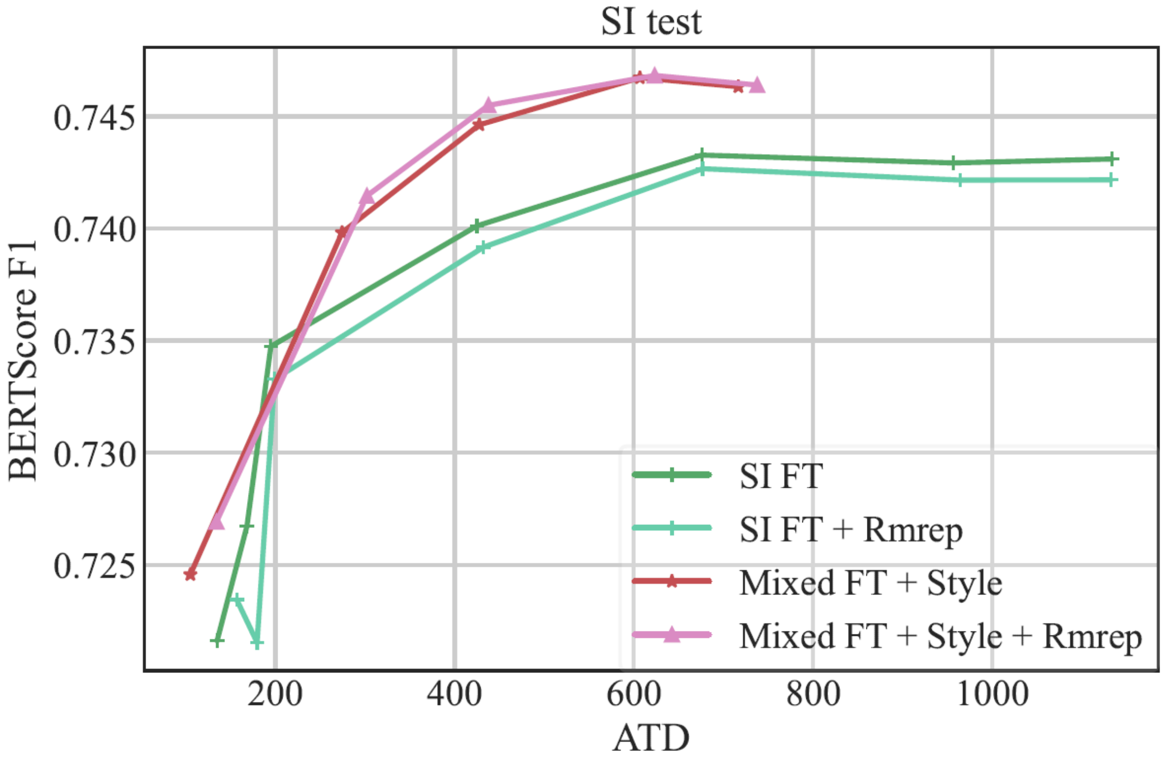


Repetition resulted in lower BLEU

Analysis: repetition by non-speech sound event label

- ▲Mixed FT Style ↔ ▲Mixed FT Style + Rmrep
  - There is no large difference in semantic similarity score (▲ ↔ ▲)
    - ➢ Removing repetition are not affecting in the semantic similarity
      - ➢ **Repetition resulted in lower BLEU, however it doesn't effect on the content of SI-like output**





Conclusion

- Background
  - The available SI data is limited
  - The trained SimulST model tends to be overfitted to SI data
- Proposed
  - Effective fine-tuning method for SimulST using mixed data of SI-style and offline-style translations with style tags
- Results
  - In BLEURT: our models were better than baselines both on SI test and offline test
    - In BLEU: our models were lower than baseline SI FT on SI test
    - Those repetitions in proposed models were not crucial for semantic translation quality
- Future work
  - Extension to other language pairs
  - Further verification via human evaluation

Reference

[Ma+2019] STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework

[Tsiamas+2022] Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano,
José Fonollosa, and Marta R. Costa-jussà. 2022. Pre-trained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022.
In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

[Toyama+2004] Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2004. Ciair simultaneous interpretation corpus. In Proceedings of Oriental COCOSDA.

[Shimizu+2013] Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In Proceedings of IWSLT.

[Doi+2021] Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), pages 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

[Zhao+2023] Zhao, Jinming, et al. "NAIST-SIC-Aligned: Automatically-Aligned English-Japanese Simultaneous Interpretation Corpus." *arXiv preprint arXiv:2304.11766* (2023).

Reference

[Chu+2017] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

[Sennrich+2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Proœedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

[Johnson+2017] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

[Caswell+2019] Isaac Caswell, Ciprian Chelba, and David Grangier.2019. Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume1: Research Papers), pages 53–63, Florence, Italy. Association for Computational Linguistics.

[Di Gangi+2019] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

[Fukuda+2023] Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Katsuhito Sudoh, Sakriani Sakti, and Satoshi Nakamura. 2023. NAIST Simultaneous Speech Translation System for IWSLT 2023. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT2023). To appear.

[Nishikawa+2023] Yuta Nishikawa and Satoshi Nakamura. 2023. Interconnection: Effective Connection between Pretrained Encoder and Decoder for Speech Translation. In Proceedings of Interspeech 2023. To appear.

[Liu+2020] Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020a. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In Proc. Interspeech 2020, pages 3620–3624.

[Kano+2022] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2022. Simultaneous neural machine translation with prefix alignment. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 22–31, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

[Kano+2023] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Average Token Delay: A Latency Metric for Simultaneous Translation. In Proceedings of Interspeech 2023. To appear.