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

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

Abstract

Simultaneous speech translation (SimulST) translates partial speech inputs incrementally. Although the monotonic correspondence between input and output is preferable for smaller latency, it is not the case for distant language pairs such as English and Japanese. A prospective approach to this problem is to mimic simultaneous interpretation (SI) using SI data to train a SimulST model. However, the size of such SI data is limited, so the SI data should be used together with ordinary bilingual data whose translations are given in offline. In this paper, we propose an effective way to train a SimulST model using mixed data of SI and offline. The proposed method trains a single model using the mixed data with style tags that tell the model to generate SI- or offline-style outputs. Experiment results show improvements of BLEURT in different latency ranges, and our analyses revealed the proposed model generates SI-style outputs more than the baseline.

1 Introduction

Simultaneous speech translation (SimulST) is a technique to translate speech incrementally without waiting for the end of a sentence. Since SimulST should work in small latency against the input speech, monotonic translation following the word order of the source language is preferable. However, making translation monotonic is not trivial especially for distant language pairs with different word orders, such as English and Japanese. Most recent SimulST studies still use parallel corpora only with offline translations and potentially have the limitation to work in a monotonic way.

A prospective approach to this problem is to use SI data to train a SimulST model for mimicking simultaneous interpretation (SI). There are several SI data resources developed so far for English-Japanese (Toyama et al., 2004; Shimizu et al., 2013; Doi et al., 2021). Despite these efforts, SI data are still very small compared to bilingual data based on offline translations. Using such scarce SI data to fine-tune an offline translation model causes overfitting on the small SI data. Training a model using mixed data of offline and SI data is another option to mitigate the problem of data scarcity, but the simple data mixture causes confusion between the output styles of offline translation and SI.

In this paper, we propose a method to train a SimulST model using mixed data of SI and offline translation with style tags to tell the model to generate SI- or offline-style output selectively. It has the advantage of sharing two different styles in a single model and generating SI-style outputs by putting the SI-style tag in the decoding, which are leveraged by offline translation data. Experiment results using MuST-C and small SI data showed improvements of BLEURT by the proposed method over the baselines in different latency ranges. Further analyses revealed that the proposed model generates more appropriate SI-style outputs than baselines.

2 Related Work

There have been many studies on simultaneous translation for text and speech in decades (Fügen et al., 2007; Oda et al., 2014; Dalvi et al., 2018). Most recent approaches are based on deep neural networks and have evolved with the technologies of neural machine translation (NMT) (Gu et al., 2017) and neural speech recognition (ASR) (Rao et al., 2017). An important advantage of the neural SimulST methods (Ma et al., 2020b; Ren et al., 2020) is their end-to-end modeling of the whole process, which improves the efficiency compared to a cascade approach. Such an end-to-end SimulST model is trained using speech translation corpora such as MuST-C (Di Gangi et al., 2019), but these corpora are usually based on offline translation due to the lack of large-scale SI data.

For the English-Japanese language pair, there

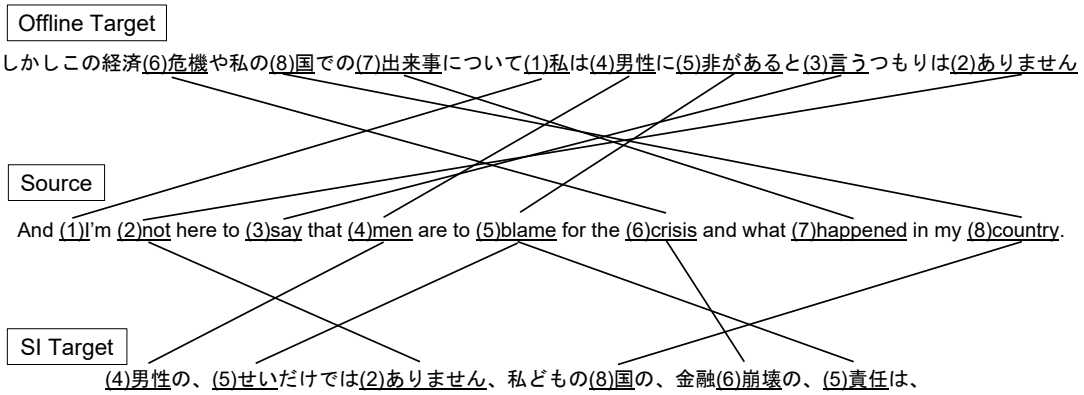


Figure 1: Example of English-to-Japanese offline translation and SI.

have been some attempts for the development of SI corpora (Toyama et al., 2004; Shimizu et al., 2013; Doi et al., 2021). However, the amount of such SI corpora is still very limited compared to offline translations. We tackle this problem by using a larger-scale offline translation corpus. This condition can be seen as domain adaptation from resource-rich offline translation to resource-poor simultaneous translation. In a typical domain adaptation scenario, an out-of-domain model is fine-tuned using in-domain data (Luong and Manning, 2015; Sennrich et al., 2016), but it tends to overfit to the small in-domain data (Chu et al., 2017). As another adaptation approach, tag-based NMT works to control the politeness of translations (Sennrich et al., 2016) and to enable zero-shot multilingual NMT (Johnson et al., 2017). This tag-based approach has been extended to multi-domain fine-tuning (Kobus et al., 2017) and mixed fine-tuning (Chu et al., 2017). These studies fine-tune NMT models using mixed data of in-domain and out-of-domain corpora. Tagged Back-Translation (Caswell et al., 2019) is an application of the tag-based approach to well-known back-translation-based data augmentation. It distinguishes source language sentences from parallel corpora and those obtained from back-translation to handle possible back-translation noise in the training of an NMT model. Our work is motivated by these tag-based methods and tackles the scarcity of SI data.

3 Differences between Offline Translation and Simultaneous Interpretation

There is a large style difference between SI and offline translation. Figure 1 shows an example of offline translation and SI transcript in Japanese for a given English source sentence. The solid lines in the figure represent word correspondences. In this figure, we can find:

- Most English content words are translated into Japanese in the offline translation, while some are missing in the SI transcript.
- The SI tries to translate the former half of the input earlier than the latter half with some unnaturalness, while the offline translation keeps naturalness in Japanese with long-distance reordering from the input English.

These points suggest important differences between offline translation and SI; SI focuses on the simultaneity of the interpretation to deliver the contents as early as possible and to maintain the interpreter's working memory. The word order difference between English and Japanese poses a serious difficulty in SI, as mentioned in the literature (Mizuno, 2017). Thus, it is important to use SI data to train a SimulST model to improve its simultaneity.

4 Proposed Method

Although training a SimulST model using SI data is necessary, we suffer from data scarcity in practice. We propose a method to use a relatively large offline translation corpus to mitigate for the SI data scarcity for training a SimulMT model. Following the tag-based NMT studies, we put a style tag at the beginning of the target string in training and predict a specified tag forcibly at the first step in inference. In this work, we use two tags: `<si>` for SI and `<off>` for offline translation.

Suppose we have an SI transcript: 私は、 買った。 ペンを、 for an English input: *I bought a*

	Offline		SI	
	#segm.	#En words	#segm.	#En words
train	328,639	5,714,360	65,008	1,120,245
dev	1,369	23,059	165	2,804
test	2,841	46,144	511	8,104

Table 1: Data sizes of offline data and SI data in the number of aligned segments.

pen. as a training example. We put the SI-style tag at the beginning of the SI transcript as follows:

<si>私は、買った。ペンを、

This string is tokenized into subwords[1]:

_<␣si␣>␣私 は␣、 ␣買 っ␣た␣。 ␣ペ ン␣を␣、

Here, we assume we have a pre-trained sequence-to-sequence model such as mBART (Liu et al., 2020b; Tang et al., 2021) as a basis of the SimulST model, as described later in the next section. The aforementioned style tags may not be included in the subword vocabulary of the pre-trained model and are tokenized further like "_<␣si␣>", but it works in practice.

5 Experimental Setup

5.1 Dataset

We used MuST-C (Di Gangi et al., 2019) v2 English-Japanese data as our offline speech translation corpus. We also prepared development and test sets from our in-house Japanese SI recordings on TED Talks that are not included in the training sets above. As for the SI data for training, we used NAIT-SIC-Aligned (Zhao et al., 2023). This SI data is constructed by applying heuristic sentence alignment to extract parallel sentence pairs using the latest version of NAIST-SIC[2] (Doi et al., 2021). From NAIST-SIC-Aligned, we selected INTRA, AUTO-DEV and AUTO-TEST as train, dev and test data, respectively. For all the SI sets, we aligned the English text segments with the corresponding audio tracks in MuST-C using an English forced-aligner Gentle[3]. Here, we excluded segments not aligned with the source speech from the aligned dataset. Table 1 shows the size of the offline and SI data.

[1]"_" is the meta-character representing white spaces in an original string by SentencePiece (Kudo and Richardson, 2018), and "␣" represents a white space in a tokenized string.

[2]https://dsc-nlp.naist.jp/data/NAIST-SIC/2022

[3]https://github.com/lowerquality/gentle

5.2 Simultaneous Speech Translation

We used our SimulST implementation based on `fairseq` (Ott et al., 2019). It followed the system architecture of the best-scored system in the IWSLT 2022 evaluation campaign (Polák et al., 2022), which used an offline ST model in the online simultaneous decoding based on Local Agreement (LA) (Liu et al., 2020a)[4].

5.2.1 Offline ST Model

We built the initial offline ST model by connecting two pre-trained models. Firstly, we used HuBERT Large as the encoder, which consists of a feature extractor trained on 60k hours of unlabeled speech data Libri-Light (Kahn et al., 2020) and a transformer encoder layer. The feature extractor is a 7-layer convolutional layer with a kernel size of (10,3,3,3,3,2,2), a stride of (5,2,2,2,2,2,2), and 512 channels, while the transformer encoder layer consists of 24 layers. Next, we used the decoder portion of mBART50, an encoder-decoder model pre-trained with 50 language pairs, as the decoder. The decoder consists of 12 layers of transformer decoders, and the embedding layer and linear projection weights are shared, with a size of 250,000. The dimension of each layer of the transformer encoder and decoder is 1024, the dimension of the feed forward network is 4096, the number of multi-heads is 16, the activation function is the ReLU function, and the normalization method is pre-layer normalization (Baevski and Auli, 2019). These two models are connected by an Inter-connection (Nishikawa and Nakamura, 2023) that weights each transformer layer of the encoder and integrates the output tensors of each layer in a weighted sum, and a length adapter (Tsiamas et al., 2022). The length adapter is a 3-layer convolutional network with 1024 channels, the stride of 2, and the activation function of GELU.

The inputs are waveforms with a 16-kHz sampling rate that are normalized to zero mean and unit variance. During training, each source audio is augmented (Kharitonov et al., 2020) with a probability of 0.8. We train the model on MuST-C (Di Gangi et al., 2019), CoVoST-2 (Wang et al., 2020), Europarl-ST (Iranzo-Sánchez et al., 2020), and TED-LIUM (Rousseau et al., 2012). We use gradient accumulation and data parallelism to achieve a batch size of approximately 32 million

[4]We also tried wait-k (Ma et al., 2019), but LA worked better than wait-k in our pilot test.

tokens. We use Adam with $\beta_1 = 0.99$, $\beta_2 = 0.98$, and a base learning rate of 2.5×10^{-4}. The learning rate is controlled by a tri-stage scheduler with phases of 0.15, 0.15, and 0.70 for warm-up, hold, and decay, respectively, while the initial and final learning rate has a scale of 0.01 compared to base. We use sentence averaging and gradient clipping of 20. We apply a dropout of 0.1 before every non-frozen layer and use time masking for 10-length spans with a probability of 0.2, and channel masking for 20-length spans with a probability of 0.1 in the encoder feature extractor's output. The loss is the cross-entropy loss with label smoothing of 0.2. We call this trained model *base* model.

The *base* model was fine-tuned using the offline training and development sets (Table 1). During fine-tuning, we set the learning rate of 2.5×10^{-5}, saved models in every 1,000 updates, and adopted checkpoint averaging over five-best checkpoints according to the loss on the development set. We call this fine-tuned model *base+O* model. About those *base* and *base+O* models, we use the NAIST IWSLT 2023 Simultaneous speech-to-speech model for the Simultaneous Speech Translation task (Fukuda et al., 2023). We further fine-tune the *base+O* model using the SI data in the same manner to derive *base+O+S* model. Here, following (Tsiamas et al., 2022), to avoid overfitting the small SI data, the parameters of the following components were kept fixed: the feature extractor and feedforward layers of the encoder and the embedding, self-attention, and feedforward layers of the decoder.

5.2.2 Fine-tuning using Prefix Alignment

For further fine-tuning toward SimulST, we extracted prefix-to-prefix translation pairs from the available training sets using Prefix Alignment (PA) (Kano et al., 2022). PA uses an offline translation model to find prefix-to-prefix translation pairs that can be obtained as intermediate translation results using a given offline translation model. Finally, we fine-tuned the *base+O* model using the prefix pairs.

5.2.3 Compared Methods

We compared the following conditions on the final fine-tuning data:

Offline FT Fine-tuned using the prefix pairs from the offline data (baseline in offline).

(BLEURT)	SI	Offline
Offline FT	0.386	0.518
SI FT	0.359	0.347
Mixed FT	0.393	0.483
Mixed FT + Style	**0.445**	**0.522**
Mixed FT + Style + Up	0.443	0.516

Table 2: BLEURT in full-sentence offline ST on SI and offline test sets.

(BLEU)	SI	Offline
Offline FT	7.8	**16.0**
SI FT	10.9	6.3
Mixed FT	9.4	13.3
Mixed FT + Style	10.3	15.4
Mixed FT + Style + Up	**12.2**	14.2

Table 3: BLEU in full-sentence offline ST on SI and offline test sets.

SI FT Fine-tuned using the prefix pairs from the SI data (baseline in SI).

Mixed FT Fine-tuned using prefix pairs from both of the offline and SI data (baseline in mixed).

Mixed FT + Style Fine-tuned using prefix pairs from both of the offline and SI data with the style tags (proposed method).

Mixed FT + Style + Up The SI portions were upsampled in **Mixed FT + Style** to balance the data size between the offline and SI data (proposed method).

Here, the prefix pairs from the offline data were obtained using *base+O* model, and those from the SI data were obtained using the *base+O+S* model. The hyperparameter settings for the fine-tuning were the same as that for the *base+O* model.

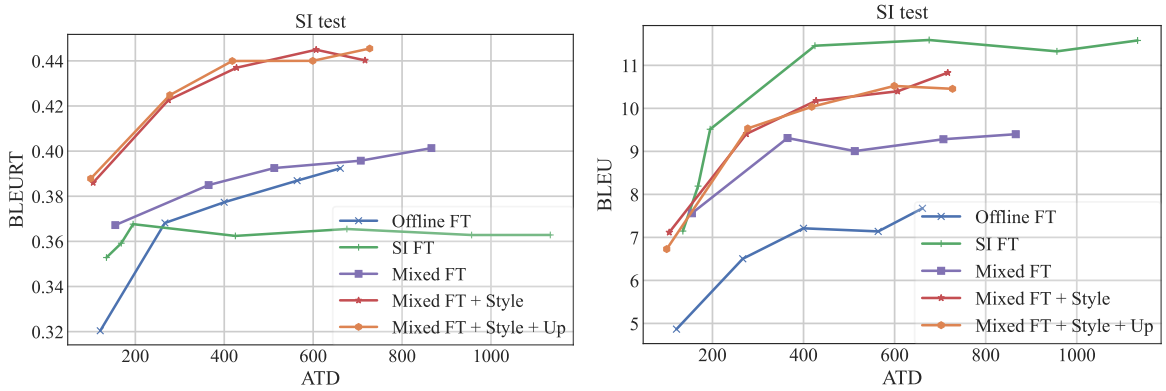
5.3 Evaluation Metrics

We evaluated the SimulST systems using SimulEval[5] (Ma et al., 2020a). The unit length of speech segments was set to {200, 400, 600, 800, 1,000} milliseconds[6]. For the SimulST systems, translation quality was evaluated in BLEURT (Sellam et al., 2020) and BLEU (Papineni et al., 2002)[7].

[5] https://github.com/facebookresearch/SimulEval

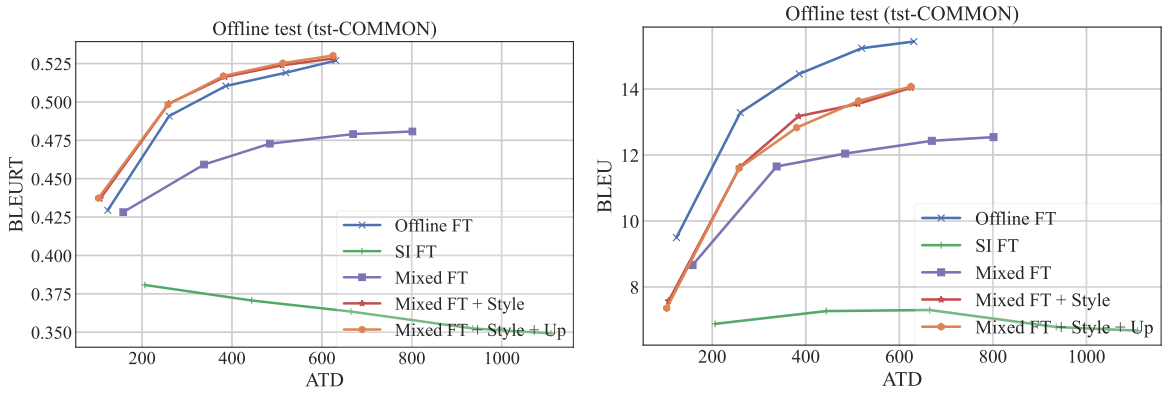
[6] We also evaluated SI FT on the SI test set with 120 and 160 ms speech segments to investigate its performance in low latency ranges.

[7] BLEU was calculated using SacreBLEU (Post, 2018).



(a) BLEURT (b) BLEU

Figure 2: SimulST latency (ATD) – quality results on SI test set.



(a) BLEURT (b) BLEU

Figure 3: SimulST latency (ATD) – quality results on offline test set.

The latency in SimulST was evaluated in Average Token Delay (ATD) (Kano et al., 2023) implemented in SimulEval. Even though Average Lagging (AL) (Ma et al., 2019) is the most popular latency metric, it sometimes resulted in negative values, as suggested by Kano et al. (2023). Thus, we present the results using ATD and include the AL results in Appendix A.

6 Results

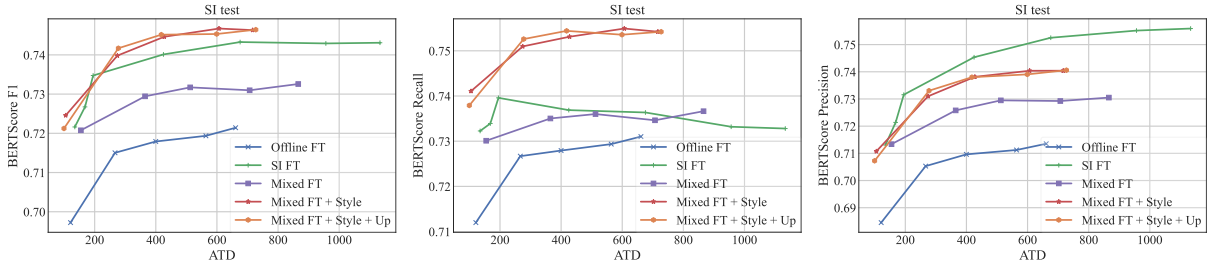
6.1 Offline Translation Results

Tables 2 and 3 show the offline translation results in BLEURT and BLEU for the SI and offline test sets. These results show that our proposed Mixed FT + Style and Mixed FT + Style + Up surpassed baselines in BLEURT for SI test. On the offline test set (MuST-C tst-COMMON), the performance of the proposed models was almost the same as Offline FT. This suggests that our proposed method leads to outputs semantically close to SI references than the baseline. Contrary, the SI FT baseline surpassed the Mixed FT + Style in BLEU. The result shows that the upsampling worked for BLEU improvement for the SI test set in the offline translation condition.

6.2 Simultaneous Translation Results

Figure 2 shows SimulST results in BLEURT and BLEU for the SI test set. In Figure 2a, the proposed method with the style tags showed clearly better BLEURT results than the baselines. The upsampling did not bring clear differences, the same as findings on the offline translation results shown in Table 2. In contrast, Figure 2b shows SI FT worked the best in almost all latency ranges, while the proposed method outperformed the other two baselines (Offline and Mixed).

Figure 3 shows SimulST results for the offline test set. They reflect the difference in reference translations between the SI and offline test sets. The Offline FT baseline worked well in BLEURT and outperformed the proposed method in BLEU. The other baselines resulted in worse BLEURT and BLEU scores than the proposed method.



(a) BERTScore F1 (b) BERTScore Recall (c) BERTScore Precision

Figure 4: SimulST latency (ATD) – quality (BERTScore) results on SI test set.

These results suggest the proposed method conveys the information given in source language speech better than the baselines.

7 Discussions

The results shown in Figures 2, 3 demonstrated the advantage of the proposed method in BLEURT, but not in BLEU. In this section, we discuss the results in detail to reveal which model works the best from the viewpoint of SimulST.

7.1 BERTScore Details

Figure 4 shows the detailed results in F1, recall, and precision by BERTScore (Zhang et al., 2020) for the SI test set. The proposed method worked the best in BERTScore recall, and the recall curves look similar to BLEURT curves shown in Figure 2a. On the other hand, the SI FT baseline worked the best in BERTScore precision, and the precision curves look very similar to the BLEU curves shown in Figure 2b. We conducted further analyses below to investigate the mixed results in different quality metrics.

7.2 Length Differences

First, we focus on the length differences between translation outputs and references. Figure 5 shows the length ratios of translation results and their references. The proposed method resulted in longer outputs than the baselines, and the SI FT baseline preferred shorter output than the others and references. From the viewpoint of the precision of the translation results, outputs longer than their references are unfavorable. Figure 6 shows the histogram of length differences between SI FT and Mixed FT + Style. They showed different distributions; this suggests that SI FT suffered from under-translation, and the proposed method suffered from over-translation.

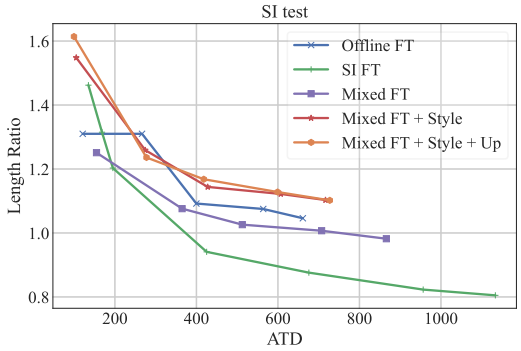


Figure 5: Length ratio results on SI test set.

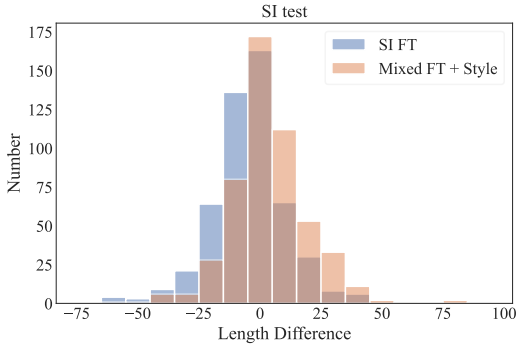


Figure 6: The length differences between hypotheses and references in SI FT and Mixed FT + Style (speech segment size is 600ms) on SI test set.

Table 4 shows the translation examples by SI FT and Mixed FT + Style. Here, SI FT generates very short outputs compared with Mixed FT + Style; BLEU is not always good due to the brevity penalty, but SI FT would have an advantage in BERTScore precision.

7.3 Non-speech Sound Events and Repetitions

Next, we investigated the over-translation suggested in the analyses above.

We observed serious repetitions by the proposed method, such as (拍手) (拍手) ..., which means (Applause). This kind of non-speech sound events (applause and laughter) are found many times in

Source	TEMPT was one of the foremost graffiti artists in the 80s. There's no hospital that can say "No." 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*) 麻痺した人達は、 (*paralyzed people*)
Mixed FT + Style (Proposed)	テンプトは、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 reference	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。 ((*He) was a great graphic artist in the 80s.*) 病院も、ノーとは言えない。(*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。 (*Everybody who is paralyzed can use this.*)
Offline reference	80年代を代表するグラフィティ・アーティストでした 病院もダメと言えません 全身麻痺の人誰もが目だけで絵を描いたりコミュニケーションできます

Table 4: Example sentences in SI FT and Mixed FT + Style (speech segment size: 600ms) on SI test set.

TED Talks, but they are not translated by interpreters and excluded from the SI data. According to this assumption, we tried to eliminate typical repetitions as follows and to conduct the evaluation after that.

- Removing tokens if they are surrounded by "()" and "<>". (if the tokens include parts of "(拍手)" like "拍手)" or "(", they were also excluded.)

- Stopping the generating output when at least one kind of 3-gram appeared at least 3 times in the steps until reaching the end of the sentence.

We applied this repetition removal on the results by Mixed FT + Style and SI + Style; they are labeled as Mixed FT + Style + Rmrep and SI FT + Rmrep, respectively. Figure 7 shows BLEU and length ratio results before and after the repetition removal. BLEU increased consistently on the proposed method while almost no changes were observed on the SI FT baseline except for one sample at ATD=200. This suggests the existence of many repetitions in the translation results by the proposed method. We also investigated BLEURT and BERTScore, as shown in Figure 8. The repetition removal made almost no changes in BLEURT, probably due to the semantic-oriented evaluation strategy of BLEURT. BERTScore Precision and F1 of the proposed method increased in the middle latency ranges, while they decreased almost consistently for the SI FT baseline. These findings suggest an over-translation problem with the proposed method, but it made little impact on semantic-oriented automatic evaluation results.

8 Conclusion

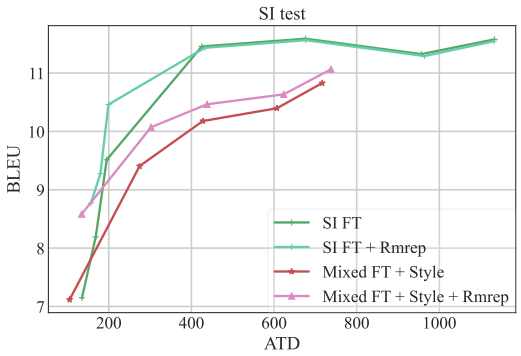
In this paper, we proposed an effective method to train a SimulST model using mixed data of SI- and offline-style translations with style tags to tell the model to generate outputs in either style, motivated by the tag-based approach to domain adaptation. Experiment results on English-to-Japanese SimulST demonstrated the advantage of the proposed method in BLEURT and BERTScore recall despite the inferior performance in BLEU and BERTScore precision due to over-translations and repetitions. Future work includes an extension to other language pairs and further verification via human evaluation.

9 Limitation

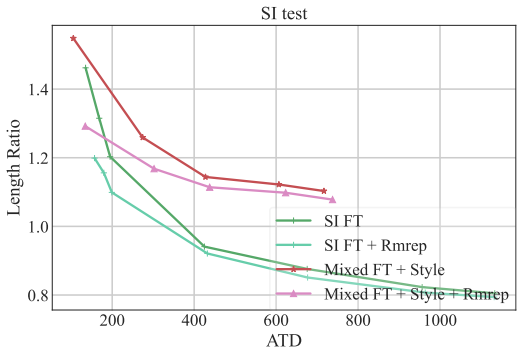
The scores reported in the SI test were lower than those in the offline test. Reporting results on other SI data would support seeing the effectiveness of our method. To our knowledge, this is the first work to use SI data as speech translation data. There are no other language pairs SI data than English-Japanese pairs those source speech and target text aligned.

Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Number JP21H05054 and JST SPRING Grant Number JPMJSP2140.



(a) BLEU



(b) Length ratio

Figure 7: Results with repetition removal (Rmrep) in BLEU and length ratio against ATD on SI test set.

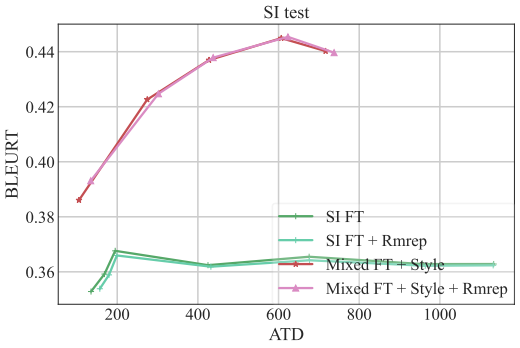
References

Alexei Baevski and Michael Auli. 2019. Adaptive Input Representations for Neural Language Modeling. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

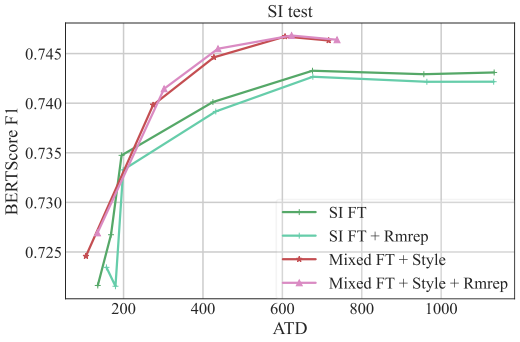
Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

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.

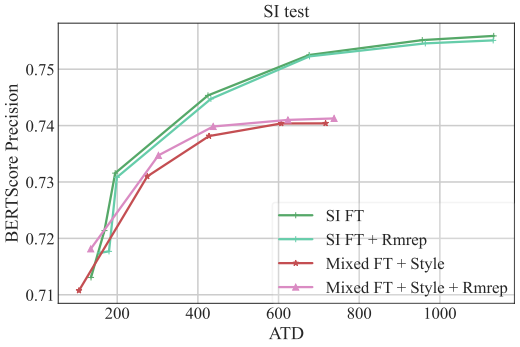
Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.



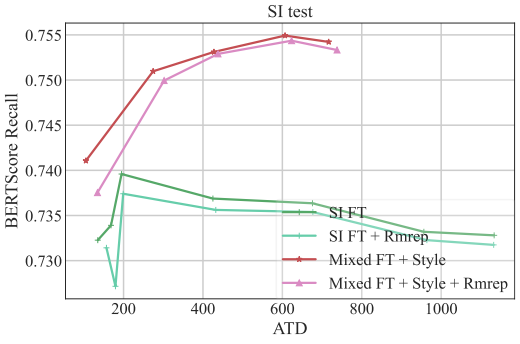
(a) BLEURT



(b) BERTScore-F1



(c) BERTScore-Precision



(d) BERTScore-Recall

Figure 8: Results with repetition removal (Rmrep) in BLEURT and BERTScore F1, precision and recall against ATD on SI test set.

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.

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.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21:209–252.

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.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

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.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. https://github.com/facebookresearch/libri-light.

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.

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

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2020. Data Augmenting Contrastive Learning of Speech Representations in the Time Domain. *arXiv preprint arXiv:2007.00991*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.

Akira Mizuno. 2017. Simultaneous interpreting and cognitive constraints. *Bull. Coll. Lit*, 58:1–28.

Yuta Nishikawa and Satoshi Nakamura. 2023. Interconnection: Effective Connection between Pre-trained Encoder and Decoder for Speech Translation. In *Proceedings of Interspeech 2023*. To appear.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 551–556.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online. Association for Computational Linguistics.

Anthony Rousseau, Paul Deléglise, and Y. Estève. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In *International Conference on Language Resources and Evaluation*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings 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.

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*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2004. CIAIR Simultaneous Interpretation Corpus. In *Proceedings of Oriental COCOSDA*.

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.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages

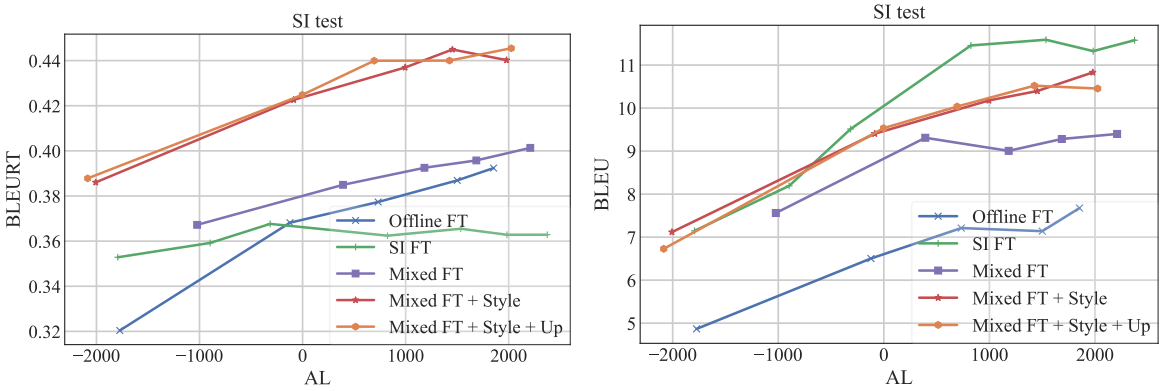
4197–4203, Marseille, France. European Language Resources Association.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinming Zhao, Yuka Ko, Ryo Fukuda, Katsuhito Sudoh, Satoshi Nakamura, et al. 2023. NAIST-SIC-Aligned: Automatically-Aligned English-Japanese Simultaneous Interpretation Corpus. *arXiv preprint arXiv:2304.11766*.

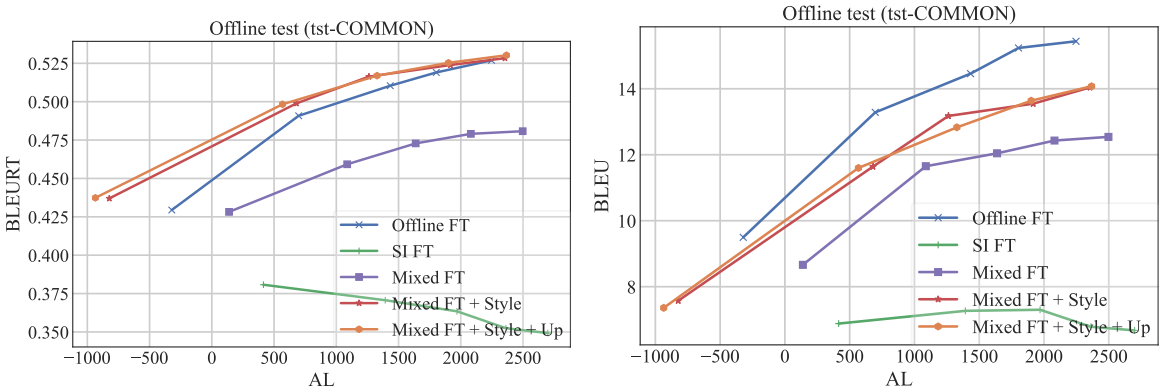
A Evaluation Results in AL.

Figure 9 shows the main results in BLEURT and BLEU in SI test in AL. Figure 10 shows the main results in BLEURT and BLEU in offline test in AL. Those results trends are almost the same as the trends in main results in Figure 2, 3.



(a) BLEURT (b) BLEU

Figure 9: SimulST latency (AL) – quality results on SI test set.



(a) BLEURT (b) BLEU

Figure 10: SimulST latency (AL) – quality results on offline test set.