General Paper Peer-Reviewed

End-to-end Simultaneous Speech Translation with Style Tags using Human Simultaneous Interpretation Data

Yuka Ko[a], Ryo Fukuda[a], Yuta Nishikawa[a], Yasumasa Kano[a],
Katsuhito Sudoh[a,b], Sakriani Sakti[a] and Satoshi Nakamura[a,c]

Simultaneous speech translation (SimulST) translates speech incrementally, requiring a monotonic input-output correspondence to reduce latency. This is particularly challenging for distant language pairs, such as English and Japanese, as most SimulST models are trained using offline speech translation (ST) data, where the entire speech input is observed during translation. In simultaneous interpretation (SI), a simultaneous interpreter translates source language speech into target language speech without waiting for the speaker to finish speaking. Therefore, the SimulST model can learn SI-style translations using SI data. However, owing to the limited availability of SI data, fine-tuning an offline ST model using SI data may result in overfitting. To address this problem, we propose an efficient training method for the speech-to-text SimulST model using a combination of small SI and relatively large offline ST data. We trained a single model with mixed data by incorporating style tags to instruct the model to generate either SI or offline-style outputs. This approach, called *mixed fine-tuning with style tags*, can be extended further using the *multistage self-training* approach. In this case, we use the trained model to generate pseudo-SI data. Our experimental results for several test sets demonstrated that our models trained using mixed fine-tuning and multistage self-training outperformed baselines across various latency ranges.

Key Words: *Simultaneous Speech Translation, Simultaneous Interpretation, Domain Adaptation*

1 Introduction

Simultaneous speech translation (SimulST) incrementally translates speech without waiting for a sentence to end. As SimulST aims for a monotonic translation process to keep up with the

[a] Nara Institute of Science and Technology
[b] Nara Women's University
[c] The Chinese University of Hong Kong, Shenzhen

Part of this work was presented at the 20th International Conference on Spoken Language Translation (Ko et al. 2023).

input speech, it is essential to maintain the original word order of the source language. However, such monotonic translation is challenging for distant language pairs with different word orders, such as English and Japanese.[1] This is because most recent SimulST studies have relied solely on parallel corpora with offline translations. Offline speech translation (ST) prioritizes fluency and accuracy over speed, resulting in the long-distance reordering of source words. Consequently, training SimulST systems with offline ST data leads to difficulties in producing monotonic translations. By contrast, simultaneous human interpreters prioritize generating translations as quickly as possible while maintaining monotonicity between the source and target content. By utilizing sentence-level human simultaneous interpretation (SI) data, in which the source and target contents correspond monotonically, it is possible to train SimulST models to mimic the monotonic translation style of human interpreters. However, recent SimulST systems cannot learn such SI-like translations because they use offline ST data.

To address this problem, SI data is used to train a SimulST model to effectively mimic SI. Over the past few decades, several English-Japanese SI corpora have been developed (Toyama et al. 2004; Shimizu et al. 2013; Matsushita et al. 2020; Doi et al. 2021). Despite these efforts, the amount of SI data remains significantly smaller than that of bilingual data based on offline translations. Fine-tuning an offline ST model with limited SI data can lead to overfitting. Another approach involves training the model using offline and SI data to address data scarcity. However, a simple mixture of data can confuse the different output styles between the offline ST and SI.

In this paper, we propose a method for training a SimulST model in a speech-to-text setting using a combination of SI and offline ST data that incorporate style tags to instruct the model to generate either SI-style or offline-style outputs. We call this primary approach *mixed fine-tuning with style tags*. The proposed method enables SI-style output using SI-style tags during decoding. In addition, leveraging the rich linguistic patterns and fluent translation capabilities learned from large-scale offline ST data helps the model differentiate between two output styles when guided by appropriate style tags. This approach can be further extended by a *multistage self-training* approach using a trained model to generate pseudo-SI data. We used pseudo-SI data to fine-tune the trained model, and this process was repeated. The motivation for this approach is to alleviate SI data scarcity using pseudo-SI data.

Experimental results demonstrated that our *mixed fine-tuning with style tags* approach improves the BLEURT (Sellam et al. 2020) and COMET-QE (Rei et al. 2020, 2021) scores compared to the baselines across various latency ranges in three English-Japanese test sets: (1) offline test

[1] This is because English is an SVO language, while Japanese is an SOV language.

of tst-COMMON in MuST-C v2 (Di Gangi et al. 2019), (2) SI reference test of NAIST-SIC-Aligned-ST (Ko et al. 2023), and (3) chunk-wise monotonic translation (CMT) reference test (Fukuda et al. 2024). We also observed further improvements when the proposed *multistage self-training* approach was applied. Further analyses showed that the proposed models produced more SI-style outputs than baseline models in terms of semantic sentence similarity in various test sets.

2 Related Work

Over the past few decades, several studies have been conducted on the simultaneous translation of both text and speech (Fügen et al. 2007; Oda et al. 2014; Dalvi et al. 2018). The most recent approaches are based on deep neural networks and have evolved with advancements in neural machine translation (NMT) (Gu et al. 2017) and neural automatic speech recognition (Rao et al. 2017). A key advantage of neural SimulST methods (Ma et al. 2020b; Ren et al. 2020) is their end-to-end modeling of the entire process, which enhances the efficiency compared to cascade approaches. These end-to-end SimulST models are typically trained using offline ST corpora, such as MuST-C (Di Gangi et al. 2019), which comprises subtitles from TED talks.[2] When training on offline data, the model was set up to perform full-sentence translation, where it waits until the entire speech input is received before starting the translation process. Offline translation tends to be natural and fluent, but is not translated with less latency. Although offline ST data are not designed for learning SimulST, they have been mainly used for training SimulST because large-scale offline data can be easily prepared.

For the English-Japanese language pair, efforts have been made to develop SI corpora (Toyama et al. 2004; Shimizu et al. 2013; Matsushita et al. 2020; Doi et al. 2021). However, the number of SI corpora is limited because of the challenges in developing high-quality SI data. In this study, we address the shortage of SI data by utilizing not only SI data but also large-scale offline translation data. One possible method is to transfer the translation skills learned from a model trained on large offline ST data (for example, vocabulary, syntactic structures, and common linguistic patterns) to the SimulST task. This can be considered as a domain adaptation from resource-rich offline translations to simultaneous resource-poor translations. In typical domain adaptation, an out-of-domain model is fine-tuned with in-domain data (Luong and Manning 2015; Sennrich et al. 2016); however, this often leads to overfitting owing to the small size of the in-

[2] `http://www.ted.com`

domain data (Chu et al. 2017). An alternative approach is the tag-based NMT, which has been used to control translation politeness (Sennrich et al. 2016) and enable zero-shot multilingual NMT (Johnson et al. 2017). This tag-based method has been extended to multidomain fine-tuning (Kobus et al. 2017) and mixed fine-tuning (Chu et al. 2017), where NMT models are fine-tuned using a mix of in-domain and out-of-domain data. Tagged back translation (Caswell et al. 2019) applies this approach to back-translation-based data augmentation, distinguishing between source-language sentences from parallel corpora and those with potential noise obtained from back-translated data. Our study was inspired by tag-based domain-adaptation methods to address the scarcity of SI data. For domain adaptation from resource-rich offline translation to resource-poor simultaneous translation while mitigating overfitting for SI data, we focused on controlling style differences between SI-style and offline-style outputs using tagging approaches.

3 Differences between Offline Translation and Simultaneous Interpretation

Because SI is a cognitively demanding task, simultaneous human interpreters employ strategies such as segmentation, summarization, and generalization (He et al. 2016). In practice, they often follow the speaker's speech using summarizations or omissions, and maintaining the word order of the source language is a crucial strategy to minimize delays and alleviate cognitive load. This approach is particularly important for distant language pairs with different word orders such as English and Japanese (Mizuno 2017). These SI-specific strategies (for example, summarization, omission, and strict maintenance of word order) result in SI outputs that are stylistically different from the offline translation outputs, thereby highlighting the significant style differences between SI and offline translation.

Figure 1 illustrates this difference using an example of a Japanese offline translation and an SI transcript for a given English source sentence. The solid lines in the figure indicate word correspondence. From this figure, the following can be observed:

- Most English words are translated into Japanese in the offline translation, whereas some of these words are missing in the SI transcript.
- The SI attempts to translate the first half of the input earlier than the second half, resulting in some unnaturalness. By contrast, offline translation maintains naturalness in Japanese by reordering the input English over longer distances.

These observations highlight the key differences between offline translation and SI. SI prioritizes simultaneity in delivering content as quickly as possible and manage the interpreter's working memory. Therefore, it is crucial to use SI data to train the SimulST model to enhance its simultaneity.

4 Proposed Method

4.1 Mixed Fine-tuning with Style Tags

We propose a method that utilizes a relatively large offline translation corpus to mitigate the lack of SI data for training the SimulST model. Figure 2 presents an overview of *mixed fine-tuning with style tags*. Using this method, we train a single model using mixed offline and SI data. In addition, when training the SimulST model, we incorporated style tags to instruct the model to generate either SI or offline-style outputs. The proposed method enables the distinction of both output styles with style tags while utilizing the knowledge of large amounts of offline data

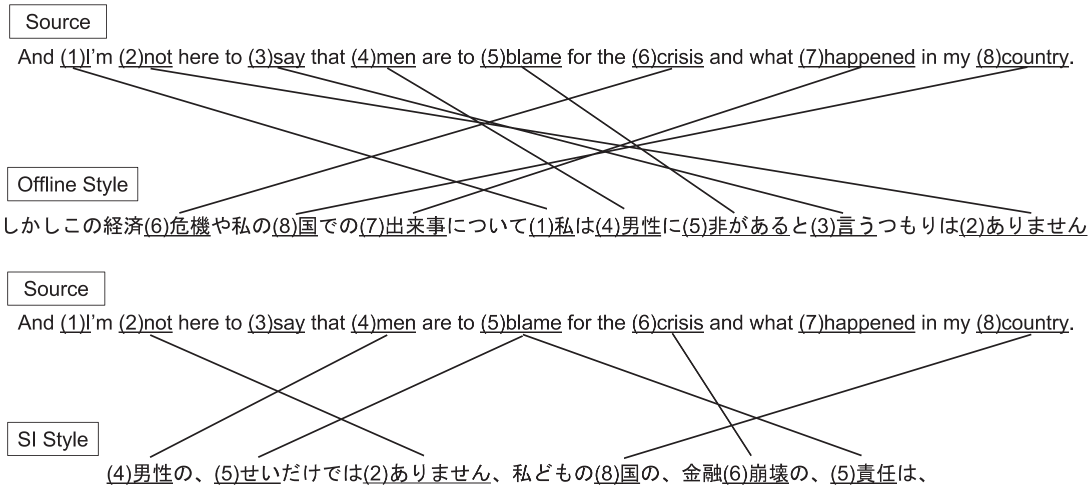

Source
And (1)I'm (2)not here to (3)say that (4)men are to (5)blame for the (6)crisis and what (7)happened in my (8)country.
Offline Style
しかしこの経済(6)危機や私の(8)国での(7)出来事について(1)私は(4)男性に(5)非があると(3)言うつもりは(2)ありません
Source
And (1)I'm (2)not here to (3)say that (4)men are to (5)blame for the (6)crisis and what (7)happened in my (8)country.
SI Style
(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、

Figure 1 Example of English-to-Japanese offline translation and SI. The example is from NAIST-SIC-Aligned (Zhao et al. 2024).

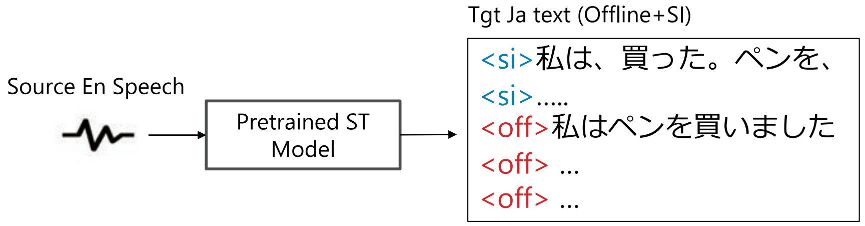


Figure 2 Mixed fine-tuning with style tags

for both style outputs. It was inspired by tag-based domain-adaptation methods. This can be considered as domain adaptation from a resource-rich offline translation to a resource-poor SI.

According to a previous study by Chu et al. (2017), fine-tuning an out-of-domain model with a small amount of in-domain data often leads to overfitting. They addressed this problem by proposing a mixed training approach that leveraged both small in-domain data and large out-of-domain data using in-domain and out-of-domain tags. Our study was inspired by previous tag-based approaches that aimed to address the scarcity of SI data. Using the proposed approach, we attempted to alleviate the problem of overfitting the SimulST model to small amounts of SI data. In the proposed approach, we fine-tune the pretrained ST model using a mix of in-domain and out-of-domain offline ST data, incorporating style tags to instruct the model to produce either SI-style or offline-style outputs. This approach allows the model to leverage a large amount of out-of-domain data to mitigate overfitting while maintaining the ability to generate both styles of output. By contrast, the conventional SI fine-tuning (SI FT) approach fine-tunes the pretrained ST model using only in-domain SI data without style tags. Although straightforward, this method risks overfitting owing to the limited size of the in-domain data. In the implementation, we used style tags at the beginning of the target strings during training and enforced the prediction of a specified tag during inference. We use `<si>` tag for SI and `<off>` tag for offline translation. Assuming an SI transcript written in Japanese: 私は、買った。ペンを、and an Offline transcript: 私はペンを買いました for the English input: *I bought a pen.* as a training example; we placed the SI-style tag at the beginning of the SI transcript, as shown in the following sentence:

`<si>`私は、買った。ペンを、(“I” “bought” “a pen”)

By contrast, we placed the offline-style tag at the beginning of the offline transcript in the following sentence:

`<off>`私はペンを買いました

In the training step, we fine-tuned the pretrained ST model by training SimulST using mixed data with style tags. In the inference step, forced decoding was applied using prefix tags. When SI-style outputs were obtained, we placed `<si>` tags, and when offline-style outputs were obtained, we placed `<off>` tags. The proposed method simplifies the process by generating multiple-style outputs using a single model by applying prefix-style tags. This is significantly different from the previous straightforward approach, which fine-tuned each pretrained model directly using the data in each style.

4.2 Multistage Self-training in Mixed Fine-tuning with Style Tags

Mixed fine-tuning with style tags in Section 4.1 uses only the original SI data. The fact that the portion of the original SI data is small compared to the offline data is a concern. We believe that increasing the SI data portion by training *mixed fine-tuning with style tags* using pseudo-SI data will further alleviate the small SI data problem.

Data augmentation based on generated synthetic data is one of the main approaches Zhang and Zong (2016) proposed to fully use the source language monolingual data for NMT by self-learning with synthetic data generated from the initial NMT system and multi-task learning using two NMTs. Self-training (Zoph et al. 2020) can improve a model by generating synthetic pseudo-labeled data from unlabeled data using an initial model when labeled data are limited.

Motivated by these studies, we propose combining a self-training method, in which pseudo-generated SI data are used to learn in multistage, with the mixed fine-tuning above. Figure 3 presents an overview of the use of self-training in mixed fine tuning using style tags. The stage of multistage self-training is denoted by N. In the $N = 2$ stage, using a pretrained SimulST, we generate pseudo-SI data by applying `<si>` tags to source speech inputs in offline ST training data. Then, we fine-tune the pretrained SimulST with the generated pseudo-SI data with `<si>` tags with original SI and offline data in our *mixed fine-tuning with style tags* approach. Because we used source speech inputs from offline ST training data to generate pseudo-SI data, the amount of pseudo-SI data was as large as that of the offline ST data. The pseudo-SI data were used for training with *mixed fine-tuning with style tags* in addition to the original SI and offline ST data. We refer to the newly trained model as Style-MultiAug-2 FT in Figure 3. This operation continues until $N_{\max}$ and the final Style-MultiAug-$N_{\max}$ FT model is adopted as the proposed model. Thus, pseudo-SI data can be generated using the models already trained in the previous step. Using this approach, we expect that when N increases, SimulST is trained using more

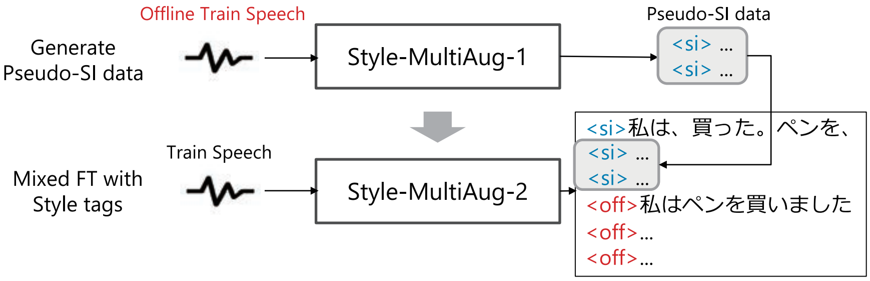


Figure 3 Multistage self-training in mixed fine-tuning with style tags

diverse pseudo-SI data. In this method, the SimulST model was trained using only self-generated data from previously trained models. Therefore, the preparation of additional ST data or external resources for preparing the pseudo-SI data are not required.

5 Experimental Setup

The following experiments investigated the effectiveness of the *mixed fine-tuning with style tags* and *multistage self-training* approaches with several test data in full-sentence offline ST and SimulST tasks, compared to baseline approaches without style tags.

5.1 Dataset

For the SI training data, we used NAIST-SIC-Aligned-ST (Ko et al. 2023).[3] The SI data were English-Japanese speech-to-text data, which were created by automatic alignment between the parallel text segments from NAIST-SIC-Aligned (Zhao et al. 2024) and the corresponding audio tracks in MuST-C. The NAIST-SIC alignment was constructed by applying automatic sentence alignment using NAIST-SIC[4] (Doi et al. 2021). For NAIST-SIC-Aligned, we selected INTRA and AUTO-DEV portions as the training and development data, respectively. Segments that did not align with the source speech were excluded from the aligned dataset. Table 1 shows the size of all data. For the offline training data, we used MuST-C (Di Gangi et al. 2019) v2 English-Japanese data as our offline speech translation corpus. The following three test sets were used for evaluation:

Offline test

The tst-COMMON portion of MuST-C (Di Gangi et al. 2019) was used as the offline test set. For the offline data, the source speech content was maintained in the target reference without omissions. While naturalness and fluency are considered in offline ST, latency is not taken into

	Offline	SI	CMT
Train	328,639	65,083	—
Dev	1,369	165	—
Test	2,841	511	511

Table 1 Data sizes of offline, SI, and CMT data for fine-tuning.

[3] https://dsc-nlp.naist.jp/data/NAIST-SIC/Aligned-ST
[4] https://dsc-nlp.naist.jp/data/NAIST-SIC/2022

account. In Japanese, the distances between the source words and the corresponding target words are generally large.

SI reference test

We used the test set in NAIST-SIC-Aligned-ST (Ko et al. 2023), which is based on the AUTO-TEST portion of NAIST-SIC-Aligned (Zhao et al. 2024). We can evaluate model performance using SI sentences, which include omissions and reordering from real professional human simultaneous interpretations.

Chunk-wise Monotonic Translation (CMT) reference test

We also used the NAIST English-to-Japanese CMT reference dataset[5] to evaluate the SimulST models. The CMT reference test data were developed by Fukuda et al. (2024) to isolate the effects of word order differences from translation omissions, a factor that often arises in simultaneous interpretation owing to summarization and generalization strategies. Both appropriate omissions and omissions resulting from interpreter errors are present in actual SI data, making it challenging to distinguish between them. Fukuda et al. (2024) also highlighted the difficulty for SI models to learn less critical phrases, advocating for SI models that focus solely on maintaining the source word order. In a related study, Doi et al. (2024) analyzed the characteristics of CMT sentences and identified grammatical structures that complicate monotonic translations in English-Japanese SI, suggesting that conventional SI reference test sets may underestimate model performance. Both studies supported the use of a monotonic translation-based dataset for a more reliable evaluation of SimulST models. Accordingly, we employed the CMT reference test set to assess the ability of SimulST to maintain the word order and preserve the speaker's original content without the confounding effects of summarization or omission.

5.2 Offline Speech Translation

Recently, one of the main approaches to SimulST has been to utilize a large-scale pretrained offline ST model with an online simultaneous decoding policy. Tsiamas et al. (2022) constructed an end-to-end multilingual offline ST model based on large pretrained speech and text models to reduce the required amount of ST data. They used wav2vec 2.0 or Hidden-Unit BERT (HuBERT) (Hsu et al. 2021) as a speech encoder, and mBART50 (Liu et al. 2020b; Tang et al. 2021) as a text decoder. Both wav2vec 2.0 and HuBERT are based on the same transformer speech encoder

[5] https://dsc-nlp.naist.jp/data/NAIST-SIC/Aligned-Chunk_Mono-EJ

architecture and are pretrained using self-supervised learning. They directly take the raw speech waveform as input, from which they first extract low-level acoustic features and then process these features to generate contextualized representations. Wav2vec 2.0, was trained to identify the true speech representation from a masked time step by solving a contrastive task on quantized representations. HuBERT predicts the masked time steps by computing the loss against pseudo-labels obtained from interactive offline clustering. mBART50 is an encoder-decoder transformer-based language model that is trained to reconstruct a sequence from its noisy version and is later extended to a multilingual version. The best ST system in Tsiamas et al. (2022) comprises HuBERT as a speech encoder and mBART50 as a text decoder. ST models pretrained using self-supervised learning can achieve highly competitive results even when a limited amount of labeled data is available. Liu et al. (2020a) proposed the local agreement (LA) approach as an online simultaneous decoding policy that outputs the agreeing prefixes of the two consecutive chunks. Polák et al. (2022) applied the LA decoding policy for an offline ST model based on pretrained wav2vec 2.0 and mBART50, resulting in the best-scored system in the IWSLT 2022 evaluation campaign.

Following these previous studies, our system is based on offline ST models using pretrained HuBERT (Hsu et al. 2021) and mBART50 (Liu et al. 2020b; Tang et al. 2021). First, we constructed an initial offline ST model by connecting these two pretrained models, as described in Section 5.2.1. We then fine-tuned the initial models using SI data in the baseline and proposed approaches, as described in Sections 5.3.1 and 5.3.2. Finally, we applied the LA simultaneous decoding policy to the trained ST models.

5.2.1 Initial offline ST model before fine-tuning with SI and offline data

Our ST models used the pretrained HuBERT (Hsu et al. 2021) and mBART50 (Liu et al. 2020b; Tang et al. 2021). Our SimulST system was implemented using `fairseq` (Ott et al. 2019) based on the code available in Tsiamas et al. (2022).[6] We used 250,000 unigram vocabularies with SentencePiece (Kudo and Richardson 2018) attached to mBART50. These two models were connected using an interconnection (Nishikawa and Nakamura 2023) that assigned weights to each transformer layer of the encoder and integrated the output tensors of each layer through a weighted sum and length adapter (Tsiamas et al. 2022). The length adapter was a three-layer convolutional network with 1,024 channels, a stride of two, and a GELU activation function. The inputs were waveforms sampled with a 16-kHz sampling rate normalized to zero mean and unit

[6] `https://github.com/mt-upc/iwslt-2022`

variance. During training, we used WavAugment (Kharitonov et al. 2020) with a probability of 0.8 for augmenting source audio. The model was trained on the 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) datasets. Gradient accumulation was applied with 24 steps and data-parallel computations to achieve a batch size of approximately 24 million tokens. The model was optimized using Adam, with $\beta_1 = 0.99$ and $\beta_2 = 0.98$. We set the learning rate to 2.5×10^{-4} and saved the models every 1,000 updates. The learning rate was managed by a tri-stage scheduler, with warm-up, hold, and decay phrases occupying 0.15, 0.15, and 0.70 of the training time. The initial and final learning rates were scaled to 0.01. Sentence averaging and gradient clipping were set to 20. A dropout rate of 0.1 was applied before every nonfrozen layer. The output of the encoder feature extractor was augmented with time masking, where the spans of ten frames were masked with a probability of 0.2, and channel masking, where the spans of 20 channels were masked with a probability of 0.1. We used cross-entropy loss with label smoothing and a 20% probability mass. We adopted checkpoints by averaging over five of the best checkpoints according to the loss in the development set.

We used this trained model for fine-tuning with the SI and offline data.[7] During fine-tuning, we set the learning rate to 2.5×10^{-5}, saved the models every 800 updates, and selected the best checkpoint based on the loss in the development set. The training process was terminated if the development loss did not decrease for four consecutive updates. We applied gradient accumulation with 16 steps and data-parallel computations to achieve a batch size of approximately four million tokens. Following Tsiamas et al. (2022), to avoid overfitting the small SI data, the parameters of the following components were fixed: the feature extractor and feed-forward layers of the encoder, and the embedding, self-attention, and feed-forward layers of the decoder.[8]

5.3 Simultaneous Speech Translation

5.3.1 Baseline models

The baseline models without style tags were as follows:

Offline FT Fine-tuned using the offline data.

SI FT Fine-tuned using the SI data.

[7] The initial offline ST model is the same as the *base* model in the following paper for the NAIST IWSLT 2023 simultaneous speech-to-speech model for simultaneous ST tasks (Fukuda et al. 2023).

[8] In Fukuda et al. (2023) and Ko et al. (2023), the models were trained with bilingual prefix alignment pairs extracted by a prefix alignment approach (Kano et al. 2022). When applying the prefix-alignment approach, we extracted prefix-to-prefix translation pairs from the available training sets, which could be obtained as intermediate translation results using a given offline translation model.However, we did not apply the prefix alignment approach in this study, because the overall scores were better without bilingual prefix pairs.

Mixed FT Fine-tuned using both the offline and SI data.

5.3.2 Proposed models

The proposed models with style tags are as follows.

Style FT Fine-tuned using both the offline and SI data with the style tags.

Style-MultiAug-N_{max} FT Multistage self-training was applied to mixed fine-tuning with style tags approach. In this experiment, multiple fine-tuning stages were repeated until convergence was achieved using the development data at each stage. As shown in Table 3, we used the best model from the fourth stage, which results in the lowest validation loss from $N = 1$ to $N = 5$. Hereafter, we refer to the proposed model created by this method as Style-MultiAug-4 FT, where $N_{max} = 4$ in Style-MultiAug-N_{max}.[9]

In this work, the style tags, that is, `<si>` and `<off>` (see Section 4.1) are not registered in the subword vocabulary attached to mBART50 model as special symbols, and thus, they should be tokenized such as "`▁<␣si␣>`" and "`▁<␣off␣>`". Here, "`▁`" 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.

5.4 Evaluation Metrics

The SimulST systems were evaluated using SimulEval[10] (Ma et al. 2020a), which automatically performs simultaneous decoding given a policy and reports several popular latency metrics. As a SimulST decoding policy, we applied LA (LA-n) (Liu et al. 2020a; Polák et al. 2022) that

Quality Metrics	Textual	Meaning	Reference	Source
BLEU	✓		✓	
BLEURT		✓	✓	
COMET-QE		✓		✓

Table 2 Quality metrics used in our experiments

N in Style-MultiAug-N FT	1	2	3	4	5
Minimum development loss	158.550	158.353	158.209	**158.135**	158.206

Table 3 Each minimum development loss of Style-MultiAug-N FT from $N = 1$ to $N = 5$.

[9] To train the first Style-MultiAug-1 initially, we used the trained Style FT model for generating pseudo-SI data and applied it to the initial offline ST model in Section 5.2.1.

[10] `https://github.com/facebookresearch/SimulEval`

displays agreeing prefixes of n consecutive chunks. We set $n = 2$ for the local agreement (LA-2), because Polák et al. (2022) reported that it achieves the best latency-quality trade-off among various LA-n strategies. Speech segments were evaluated at intervals of {200, 400, 600, 800, 1,000} ms to control for the quality-latency trade-off. Hypotheses for speech segments were generated using a beam search with a beam size of five. The translation quality was evaluated using BLEU (Papineni et al. 2002), BLEURT (Sellam et al. 2020) and COMET-QE (Rei et al. 2020, 2021). BLEU was computed directly using SimulEval, which uses SacreBLEU (Post 2018). The BLEURT and COMET-QE scores were evaluated separately using the translations generated after all the sentences were generated using SimulEval. Table 2 lists the quality metrics used in our experiments. Because SI is a cognitively demanding task and human simultaneous interpreters try to summarize or generalize the contents considering the meaning of the speaker's speech, in this study, we focused on scores based on semantic similarity measures, such as BLEURT and COMET-QE. They are based on the embedding of a sentence rather than BLEU, which is a traditional quality metric for machine translation evaluated on textual surface agreements. Compared to BLEURT and BLEU, which are reference-based evaluation metrics, we used reference-less COMET-QE to directly measure the amount of the speaker's speech contents included in the test translation.

The latency in SimulST was evaluated using the average token delay (ATD) (Kano et al. 2024), as implemented in SimulEval. ATD is a latency evaluation metric that focuses on the end timings of partial translations in simultaneous translations.[11] Note that to generate outputs from the proposed model, we applied `<off>` tags for the offline test and `<si>` tags for the SI reference and CMT reference tests.

6 Results

6.1 Full-sentence Offline Translation Results

Table 4 lists the experimental results of full-sentence offline translation in BLEURT, COMET-QE and BLEU for the Offline, SI reference and CMT reference test sets. Note that COMET-QE in the SI and COMET-QE in CMT reference tests had the same results because the same source transcripts were used in both the SI reference and CMT reference tests. Additionally, the number of source transcripts in the offline test differed from those in the SI reference and CMT reference

[11] Although average lagging (Ma et al. 2019) is the most common latency metric; it sometimes results in negative values, as noted by Kano et al. (2024).

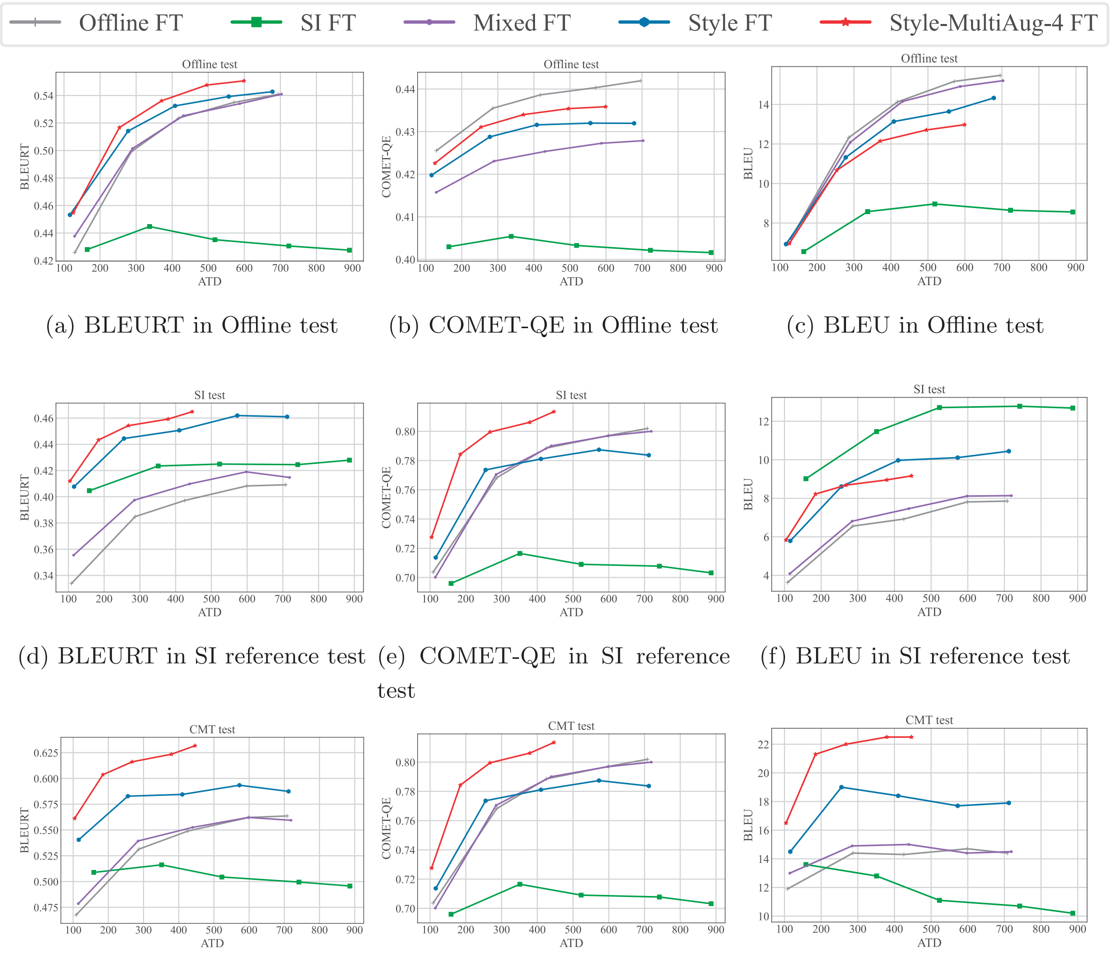
	BLEURT			COMET-QE			BLEU		
	Offline	SI	CMT	Offline	SI	CMT	Offline	SI	CMT
Baseline									
Offline FT	0.530	0.398	0.550	0.441	0.813		**16.0**	8.1	13.9
SI FT	0.417	0.420	0.480	0.409	0.688		7.8	**12.3**	8.9
Mixed FT	0.531	0.415	0.550	0.441	0.808		15.9	8.0	13.8
Proposed									
Style FT	0.539	0.458	0.579	0.445	0.778		15.4	11.1	16.6
Style-MultiAug-4 FT	**0.549**	**0.469**	**0.632**	**0.448**	**0.815**		14.4	9.5	**22.4**

Table 4 BLEURT, COMET-QE and BLEU in full-sentence offline ST on Offline, SI reference and CMT reference test sets.

tests, which resulted in different COMET-QE scores. These results show that the proposed Style FT and Style-MultiAug-4 FT surpassed the baselines in BLEURT and COMET-QE for all test sets. This finding suggests that our proposed Style FT and Style-MultiAug-4 FT can produce more semantically aligned outputs than the SI FT and Offline FT baselines in the SI reference and CMT reference tests. By contrast, the Offline FT and SI FT baselines surpassed the proposed Style FT and Style-MultiAug-4 FT in BLEU.

6.2 Simultaneous Translation Results

Figure 4 shows the quality and latency plots for SimulST in BLEURT, COMET-QE, and BLEU on the Offline, SI reference, and CMT reference test sets. Each point in each graph represents the score for speech segment sizes of 200, 400, 600, 800, and 1,000 ms from the left. In Figures 4a, 4d and 4g, the proposed Style FT and Style-MultiAug-4 FT clearly exhibit better BLEURT results than the Offline FT, SI FT, and Mixed FT baselines. Note that the COMET-QE in the SI reference test (Figure 4e) and the COMET-QE in CMT reference test (Figure 4h) had the same results because the same source transcripts were used for both the SI reference and CMT reference tests. According to the COMET-QE in the Offline test (Figure 4b), the Offline FT baseline was better than the proposed Style FT and Style-MultiAug-4 FT. From the COMET-QE in the SI reference test (Figure 4e), compared with the Offline FT baseline, the proposed Style FT exhibited low latency, and the Style-MultiAug-4 FT performed better in all latency ranges. It can be seen from the COMET-QE in the Offline and SI reference tests (Figures 4e and 4b) that both SI FT baselines were worse than the other models. These results suggest that a direct SI FT causes inappropriate omissions and fails to retain speaker content. In contrast, the Offline FT could learn to retain the speaker's content, resulting in the best COMET-QE scores in the



(a) BLEURT in Offline test (b) COMET-QE in Offline test (c) BLEU in Offline test

(d) BLEURT in SI reference test (e) COMET-QE in SI reference test (f) BLEU in SI reference test

(g) BLEURT in CMT reference test (h) COMET-QE in CMT reference test (i) BLEU in CMT reference test

Figure 4 SimulST latency (ATD) – quality results on test sets.

Offline test (Figure 4b). However, our proposed Style FT and Style-MultiAug-4 FT could also avoid excessive translation omissions compared to the SI FT baseline, even when using SI data, and they could retain the speaker's content as effectively as that in Offline FT. In particular, Style-MultiAug-4 FT demonstrates a higher tendency to achieve this than Style FT. From the overall results of BLEURT and COMET-QE, Style-MultiAug-4 FT performed better than Style FT. However, when we observed the BLEU trends in the SI reference test (Figure 4f), the baseline SI FT was the best in all latency ranges. The CMT reference test results (Figures 4g, 4h and 4i), BLEURT and BLEU trends were approximately similar. These results showed similar trends;

Style-MultiAug-4 FT and Style FT were better than the other baselines, and Style-MultiAug-4 FT was better than Style FT.

6.3 Human Evaluation by Simultaneous Interpreter

We also conducted a human evaluation by a native Japanese simultaneous professional interpreter to determine which model's output was more SI-like in the following two points.

Adequacy To check whether the translation is making inappropriate omissions. To identify the source, the speech content was retained in the translation.

Fluency To check if the translation is grammatically fluent, such as in full-sentence offline translation.

The interpreter presented three output sentences generated by the (1) baseline SI FT, (2) proposed Style FT, and (3) Style-MultiAug-4 FT models simultaneously without providing information on how the translation was generated. We asked the interpreter to assign a score of to 1-5 to the three sentences for each evaluation point: adequacy and fluency. Each set of sentences comprised 511 sentences generated by each ST model for the SI reference test evaluation. Table 5 shows the adequacy and fluency results of the baseline SI FT, proposed Style FT, and Style-MultiAug-4 FT methods evaluated by a simultaneous interpreter. These three results were obtained for a speech segment size is 400 ms. The results show that our proposed methods obtained better scores in terms of adequacy and fluency, and Style-MultiAug-4 FT obtained better scores than Style FT. The gap between Style-MultiAug-4 FT and Style FT was larger in adequacy (4.23 vs. 4.08) than in fluency (4.29 vs. 4.24). We also observed that the gap between the proposed Style-MultiAug-4 and the baseline SI FT was more significant in adequacy (4.23 vs. 3.57) compared to that in fluency (4.29 vs. 3.83). In this study, we focused on alleviating undergeneration in the baseline model, and it showed that adequacy improved significantly compared with fluency, because we did not focus directly on fluency in this study. However, we observed

	Adequacy	Fluency
SI FT	3.57 (0.972)	3.83 (0.948)
Style FT	4.08 (0.987)	4.24 (0.817)
Style-MultiAug-4 FT	4.23 (0.930)	4.29 (0.703)

Table 5 Fluency and adequacy evaluated by human simultaneous interpreter. Each set of sentences consists of all 511 sentences generated from each ST model for the SI reference test evaluation. All three results are from when the speech segment size is 400ms. The values mean averaged score and () means variance.

improvements in both fluency and adequacy. This indicates the effectiveness of our proposed methods not only for maintaining the source content, but also for maintaining the naturalness of the translation.

7 Discussions

The main results in Section 6 demonstrate the overall effectiveness of the proposed Style FT and Style-MultiAug-4 FT. However, it is not clear whether applying self-training in multiple stages is truly effective, as it remains unclear whether the selected $N = 4$ model outperforms the intermediate models in $N = 1, 2, 3$ or the models trained with upsampling approach using the original SI data. Intermediate models, such as Style-MultiAug-2 or Style-MultiAug-3, which were trained in the process of training Style-MultiAug-4, may also be sufficient. A simple upsampling approach using the original SI data may be sufficient for using pseudo-SI data in multistage self-training. By contrast, although the proposed methods were good overall, there were cases where the trend differed depending on the test set and evaluation metrics. For example, in BLEU, the baseline SI FT performed better on the SI reference test. We attribute these trend differences to variations in the outputs of the baseline and proposed methods as well as differences in the test sets. To clarify these points in detail, we conducted the following analyses:

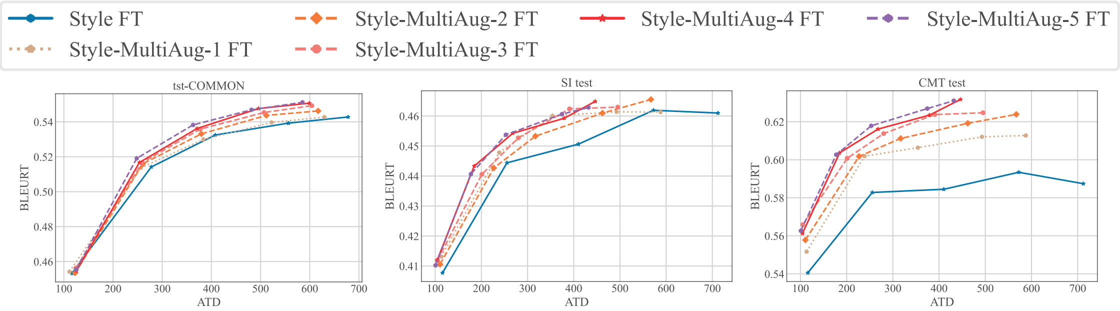
(1) We examined the score trends in intermediate models ($N = 1, 2, 3$) in multistage self-training to examine how the scores changed in each stage.

(2) We observed the results when applying upsampling to examine the effectiveness of using pseudo-SI data.

(3) We compared the output trends in the baseline and proposed methods and test sets differences to examine where the result trend differences came from.

7.1 Effectiveness of Multistage Self-training

To determine the effectiveness of multistage self-training, we compared the translation performance of the Style-MultiAug-N FT models with different N, specifically, $N = 1$ to $N = 5$. Figure 5 shows the BLEURT results of multistage self-training from Style-MultiAug-1 FT to Style-MultiAug-5 FT in the Offline, SI reference and CMT reference test sets. Among all test sets, the performance improved progressively when N increased. These results show that as the number of stages increases, the model performance improves. We expected the model to be trained with different variations in the pseudo-SI data in each step. In the Offline test (Figure 5a), although there was no significant performance improvement compared with the CMT reference



(a) BLEURT in Offline test (b) BLEURT in SI reference test (c) BLEURT in CMT reference test

Figure 5 SimulST latency (ATD) – BLEURT result in Offline, SI reference and CMT reference tests to examine the effectiveness of multistage self-training.

and SI reference tests, there was some improvement in performance. The slight improvement in the offline test can be attributed to the increasing amount of SI data as the stages progressed, whereas the amount of offline data remained unchanged. Nevertheless, we found that multistage self-training was effective in the Offline test. In this study, we selected Style-MultiAug-4 FT as the best model based on the development loss across stages from $N = 1$ to $N = 5$. Although the results from $N = 5$ showed a slight improvement over $N = 4$ in the CMT reference and Offline tests, the SI reference test results remained largely unchanged between $N = 4$ and $N = 5$. Based on the evaluation results, the selection of models based on development loss during the multistage self-training process appears to be appropriate.

7.2 Effectiveness of Upsampling

Upsampling is one of the easiest approaches for increasing the contribution of small data when training with both large and small data. We can also upsample the SI data instead of using augmented data with self-training in Style-MultiAug-4 FT. We also conducted experiments under the following settings to determine the effect of SI data upsampling.

Mixed-Up FT Fine-tuned using both the offline and SI data. The SI portions were upsampled when fine-tuned from both the offline and SI data without style tags.

Style-Up FT The SI portions were upsampled when fine-tuning using both the offline and SI data with style tags.

From the results in Table 6, we compared the Mixed FT with Mixed-Up FT or Style FT with Style-Up FT and observed that simply upsapling SI did not improve the model performance. The

	BLEURT			COMET-QE			BLEU		
	Offline	SI	CMT	Offline	SI	CMT	Offline	SI	CMT
Baseline									
Offline FT	0.530	0.398	0.550	0.441	0.813		**16.0**	8.1	13.9
SI FT	0.417	0.420	0.480	0.409	0.688		7.8	12.3	8.9
Mixed FT	0.531	0.415	0.550	0.441	0.808		15.9	8.0	13.8
Mixed-Up FT	0.510	0.420	0.536	0.433	0.778		14.0	9.4	12.5
Proposed									
Style FT	0.539	0.458	0.579	0.445	0.778		15.4	11.1	16.6
Style-MultiAug-4 FT	**0.549**	**0.469**	**0.632**	**0.448**	**0.815**		14.4	9.5	**22.4**
Style-Up FT	0.532	0.449	0.534	0.442	0.737		13.4	**12.4**	12.8

Table 6 BLEURT, COMET-QE and BLEU results in the upsampling methods in full-sentence offline ST on Offline, SI reference, and CMT reference test sets.

Mixed-Up FT was better than the Mixed FT in the SI reference test BLEU, but was worse in the Offline test BLEU as a trade-off. We also observed the same trend when comparing the Style-Up FT with the Style FT. In addition, we observed that the Style-Up FT with simple upsampling did not surpass the Style-MultiAug-4 FT. These results show that the upsampling of small amounts of SI data was ineffective for both the previous and proposed methods, which could be due to slight variations in the upsampled SI data. In contrast, the Style-MultiAug-4 FT uses diverse augmented pseudo-SI data. This would have resulted in an improvement over the Style FT and Style-Up FT.

7.3 Output trends from baseline and proposed models

To clarify the performance differences in the test sets and evaluation metrics, we first observed the output trends in the baseline and proposed models through the length differences.

7.3.1 Length differences

First, we focus on the length differences between the translation outputs and references. Figure 6 shows the length ratios between the translation results and CMT reference test. As shown in Figure 6, the proposed Style FT and Style-MultiAug-4 FT resulted in longer outputs than those of the baselines, and the SI FT baseline preferred a shorter output than the others and references. In addition, the length ratio in the SI reference test was higher than that in the CMT reference test. This is probably because SI sentences generated by human interpreters tend to contain many omissions. Moreover, because the SI FT is trained with SI data directly, it tends to be overfitted and produces significantly short outputs. In contrast, our proposed Style FT

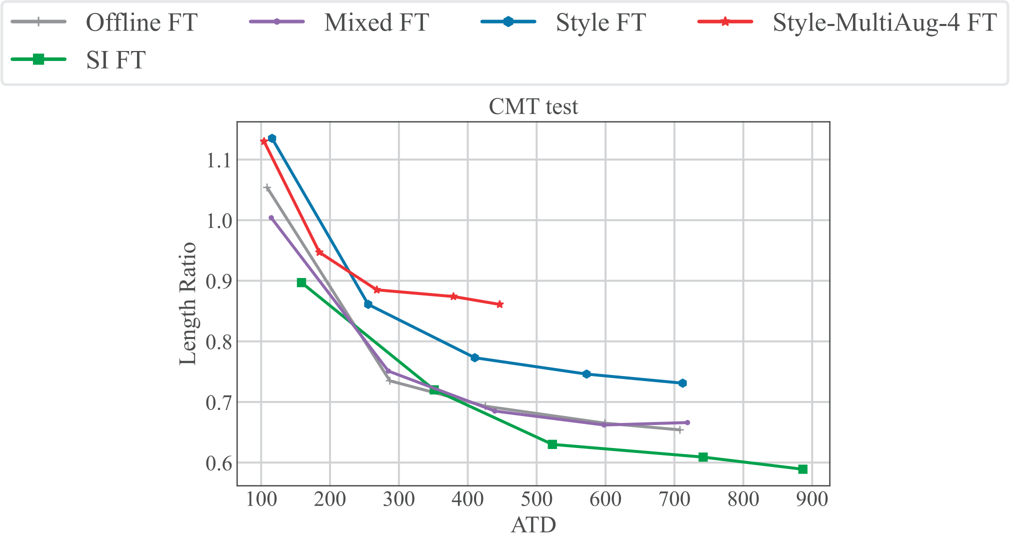


Figure 6 Length ratio results in CMT reference test.

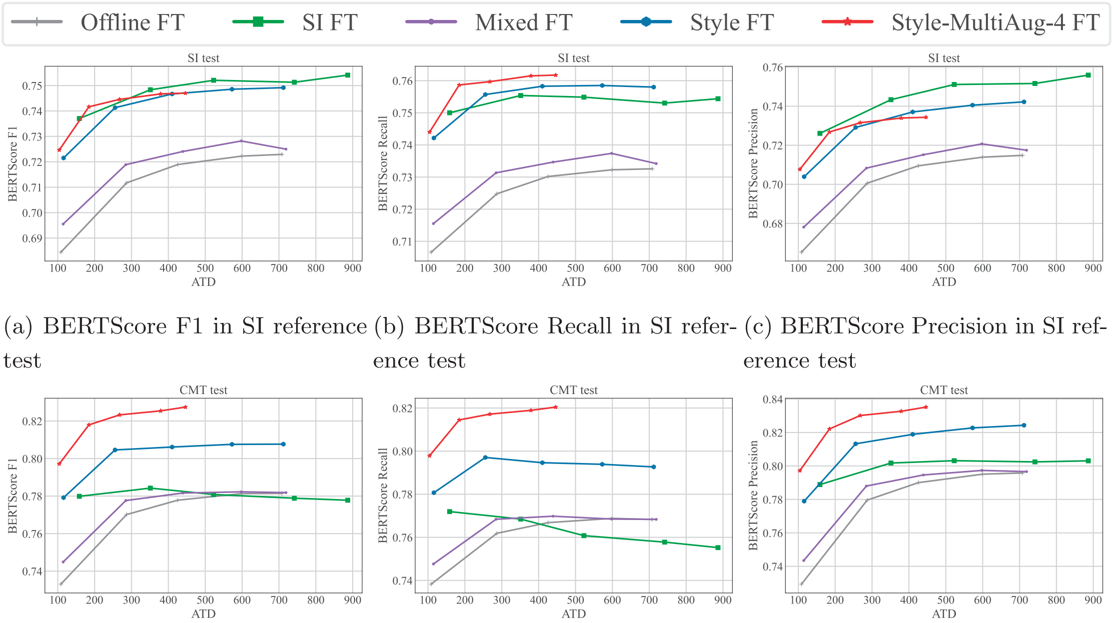
and Style-MultiAug-4 FT prefer longer outputs, alleviating the overfitting problem associated with short outputs. This suggests that the baseline SI FT suffers from undertranslation, and the proposed Style FT and Style-MultiAug-4 FT methods alleviate such undertranslation problems. However, the proposed Style FT and Style-MultiAug-4 FT methods exhibited overgeneration trends, especially in the low-latency range.

7.3.2 Undergeneration in SI FT baseline results in inflated BLEU scores in the SI reference test

We observed undergeneration trends in the baseline SI FT methods and overgeneration trends in the proposed Style FT and Style-MultiAug-4 FT methods. We believe that these output trends can lead to different results for different test sets and evaluation metrics. When we compared the trend in BLEU of the SI reference test and BLEU of the CMT reference test, the large difference was in the reference length between the SI reference and CMT reference tests. There were almost no omissions in the CMT reference test; however, there were omissions in the SI reference test. Then, in the case of short outputs from a SimulST model, even if the necessary content is included, there can be a smaller penalty for short output in BLEU. Therefore, it is difficult to evaluate SimulST for SI tasks using BLEU considering omissions. Therefore, a method to evaluate the similarity between sentences and words may be more suitable for interpretation. First, to confirm the trends in both output length and word similarity, we evaluated the BERTScore (Zhang et al.

2020) results for F1, recall, and precision.

Figure 7 shows the detailed results for F1, recall, and precision using BERTScore for the SI reference and CMT reference test sets. This indicates that the F1 scores in the proposed models were better in both the SI reference and CMT reference tests, particularly in the low-latency range. When we observed the recall, the proposed Style-MultiAug-4 FT performed the best in BERTScore recall in the SI. We expect that longer outputs in the proposed Style-MultiAug-4 FT will lead to a recall gain. By contrast, the SI FT baseline was the best for BERTScore precision in the SI reference test. The precision trend in the SI reference test (Figure 7c) was very similar to that of BLEU in the SI reference test (Figure 4f). These results suggest that BERTScore precision and BLEU scores are influenced by the output length. From the perspective of the precision of the translation results, outputs that were longer than their references were unfavorable. In addition, the SI reference test set contained several omissions. Usually, BLEU penalizes a short output via a brevity penalty, but the penalty for short hypotheses was not sufficiently applied because the references were originally short owing to omissions. Consequently, the SI FT results were overestimated by BLEU in the SI reference test.



(a) BERTScore F1 in SI reference test

(b) BERTScore Recall in SI reference test

(c) BERTScore Precision in SI reference test

(d) BERTScore F1 in CMT reference test

(e) BERTScore Recall in CMT reference test

(f) BERTScore Precision in CMT reference test

Figure 7 SimulST latency (ATD) – quality (BERTScore) results on test sets.

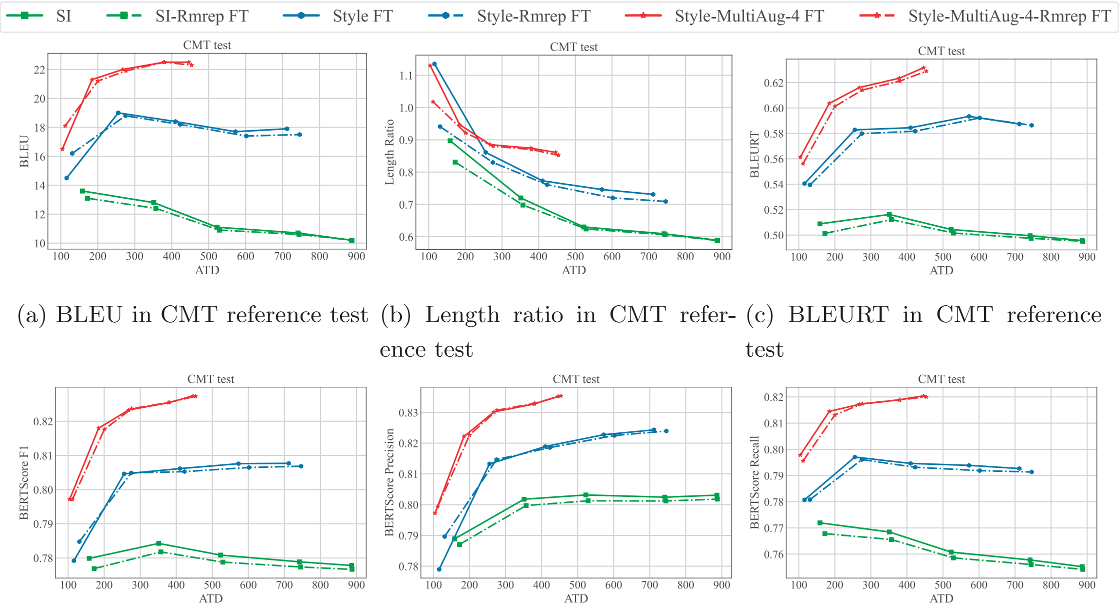
In the CMT reference test, all results showed the same trend as in BLEURT, COMET-QE, and BLEU, indicating that the proposed Style FT and Style-MultiAug-4 FT methods were better. This also indicates that the SI reference test is not suited to BLEU evaluation because it is easily affected by the under-generation effect of the baseline SI FT.

7.3.3 Non-speech events causes over-generation in proposed methods in low latency range

We observed overgeneration trends in the proposed Style FT and Style-MultiAug-4 FT methods. As shown in Figure 6, we observed overtranslation in the proposed methods, especially in the case of low latency, that is, when the speech segment size was 200 ms. We observed some repetitions by the proposed method, such as (拍手) (拍手) ..., which means (Applause). These nonspeech sound events (applause and laughter) are often found in TED Talks and are transcribed as labels in the target translation of the offline data. When a model is directly fine-tuned on SI data that contain no nonspeech event labels, the outputs of the model include few such labels. By contrast, our proposed models generated a significant number of nonspeech sound labels and repetitions related to the labels. We attribute this difference to the fact that the proposed models are trained with both mixed offline and SI data simultaneously, making them more influenced by nonspeech event labels in the offline data, even when focusing on the SI-style output with style tags. Based on this assumption, we attempted to eliminate typical repetitions as follows and then conducted an evaluation:

- Remove tokens if they are surrounded by "()" and "<>". (if the tokens include parts of "(拍手)" like "拍手)" or "(", they were also excluded.)
- Stop generating output as soon as any 3-gram appears at least three times before reaching the end of the sentence.

We applied this repetition removal to the results of the SI FT, Style FT, and Style-MultiAug-4 FT, which are labeled SI-Rmrep FT, Style-Rmrep FT, and Style-MultiAug-4-Rmrep FT, respectively. Figure 8 shows the BLEU, length ratio, BLEURT, and BERTscore scores before and after the repetition removal. We observed a larger length reduction when the speech segment size was 200 ms in the proposed Style FT and Style-MultiAug-4 FT compared with the baseline SI FT. The repetition effect was lower for the baseline SI FT than that for the proposed model. In both the proposed Style FT and Style-MultiAug-4 FT, BLEUs increased, whereas the SI FT baseline did not improve when the speech segment size was 200 ms. This suggests the existence of several repetitions in the translation results of the proposed method. In BLEURT, the baseline was lower than those of the other proposed models when the speech segment size was 200 ms. The



(a) BLEU in CMT reference test (b) Length ratio in CMT reference test (c) BLEURT in CMT reference test

(d) BERTScore F1 in CMT reference test (e) BERTScore Precision in CMT reference test (f) BERTScore Recall in CMT reference test

Figure 8 Results with repetition removal (Rmrep) in BLEU, length ratio, BLEURT and BERTScore scores in ATD on CMT reference test set.

BERTScore F1 was lower than SI FT overall compared to that for the SI-Rmrep FT; the scores between Style FT and Style-Rmrep FT or between Style-MultiAug-4 FT and Style-MultiAug-4-Rmrep FT were approximately similar, respectively. This could be due to the greater score reduction of SI-Rmrep FT and the same scores between Style FT and Style-Rmrep FT or between Style-MultiAug-4 FT and Style-MultiAug-4-Rmrep FT in BERTScore precision. Overall, these findings suggest an overtranslation trend from the repetitions caused by nonspeech event labels in the proposed methods in the low-latency range. Nevertheless, we observed no significant impact on semantic similarity, as it does not change significantly before and after excluding these repetitions.

7.3.4 Output examples

Table 7 shows example sentences from the baseline SI FT, proposed Style FT, and Style MultiAug-4 FT with 400-ms segment size. Each human evaluation score is also included. In Example 1, although the SI FT is undertranslated, the SI reference test translation is also excessively short. The SI FT was also trained directly on SI training data containing such excessively

Example 1		Ade.	Flu.
Source	It's probably the smallest of the 21 apps that the fellows wrote last year.	—	—
SI FT (Baseline)	一番小さいアプリです。*(Smallest application.)*	3	3
Style FT (Proposed)	恐らく21のアプリの中で、一番小さいものだと思います。*(It is probably the smallest of the 21 applications.)*	4	4
Style-MultiAug-4 FT (Proposed)	これは、おそらく、21のアプリの中で、最も小さいものです。昨年、フェローが書いたものです。*(This is probably the smallest of the 21 applications. It was written by a fellow last year.)*	5	4
SI reference test	昨年作ってくれたもので。*(It was made last year.)*	—	—
CMT reference test	おそらくそれは最小です、21のアプリの中で、昨年フェローが書いたものの中で。*(It is probably the smallest, out of the 21 apps the fellows wrote last year.)*	—	—
Example 2			
Source	It was running into bankruptcy last fall, because they were hacked into.	—	—
SI FT (Baseline)	破産したんです。この秋に破産したんです。*(Bankruptcy. I went bankrupt this autumn.)*	3	4
Style FT (Proposed)	これは、去年の秋に、破産したものです。なぜなら、彼らは、ハッキングされたからです。*(These are the ones that, last autumn, went bankrupt. Because they were hacked.)*	5	5
Style-MultiAug-4 FT (Proposed)	それは、昨年、破産につながったものです。なぜなら、彼らは、不正に侵入されたからです。*(It is what led to their bankruptcy last year. Because they were illegally infiltrated.)*	5	5
SI reference test	破産をしたのは、去年の秋なんです。ハッキングをされたからです、*(It was last autumn that I went bankrupt. It was because we were hacked,)*	—	—
CMT reference test	それは、昨秋、破産寸前でした、ハッキングされたためです。*(It was on the verge of bankruptcy last autumn, due to being hacked.)*	—	—

Table 7 Example sentences in SI FT, Style FT and Style-MultiAug-4 FT (speech segment size: 400ms) on SI reference and CMT reference test sets. Ade. means Adequacy and Flu. means Fluency. () is the English translation result translated by DeepL in 2024/08/06.

short SI translations, and we assume that the SI FT model was omitted excessively without understanding which parts should be omitted. However, from the source and the CMT reference test, the proposed Style FT and Style-MultiAug-4 FT retained more speaker content. The human evaluation scores of the proposed methods were higher than those of the baseline SI FT.

Example 2 also shows that the part "they were hacked into" was not translated in the baseline SI FT, indicating the under-translation problem. However, we see that the outputs of the proposed method are longer than those of the baseline SI FT, and this trend is more visible in the Style-MultiAug-4 FT than in the Style FT. However, the output tends to include more connection words and subjects, such as "これは (this)," "なぜなら (because)," and "彼らは (they)." This did not significantly affect the communication of meaning. Even if these words are removed, their meanings are still understood in many cases and do not significantly affect the performance. This suggests that the over-generated lengthy parts of the proposed method have no significant impact on the meaning of the translation.

8 Conclusion

In this study, we proposed an effective method for training a SimulST model using mixed data of SI- and offline-style translations with style tags to train the model to generate outputs in either style, motivated by the tag-based approach in domain adaptation. Experimental results on the English-to-Japanese SimulST demonstrated the overall advantage of the proposed method, particularly in BLEURT and COMET-QE in the CMT reference test set. Future work will include training a SimulST model that makes appropriate omissions, such as SI generated by human interpreters. Our proposed models retained more content in the source speech but tended to produce a longer output than the baseline models. For continuous and longer speech, a SimulST model should also be evaluated to determine whether it can keep up with the source speech. In future studies, we will also consider applying the proposed method to other language pairs. We used the NAIST-SIC-Aligned (Zhao et al. 2024) and English-Japanese SI data aligned at the utterance level to investigate the effectiveness of our proposed methods. These methods can also be applied if the utterance-level SI data are available for other language pairs. In addition, in this study, pseudo-SI data in self-training were generated from offline ST data's speech, but future work will also explore the generation of more from monolingual speech data. Recent advances in large language models (LLMs) have shown promise in translation tasks. We consider leveraging LLMs to extract distinctive characteristics from limited SI data and develop a SimulST system that properly handles appropriate omissions.

Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681 and JP24KJ1695. This paper is an extended version of that published in Proceedings of the 20th International Conference on Spoken Language Translation (Ko et al. 2023).

In this study, we expand on the ideas of previous works and conduct an additional experiment to reveal the effectiveness of our proposed multistage self-training method. We also conducted further analyses in COMET-QE on CMT reference test set to determine the effectiveness of our proposed method in maintaining source content in translations.

References

Caswell, I., Chelba, C., and Grangier, D. (2019). "Tagged Back-Translation." In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K. (Eds.), *Proceedings of the 4th Conference on Machine Translation (Volume 1: Research Papers)*, pp. 53–63, Florence, Italy. Association for Computational Linguistics.

Chu, C., Dabre, R., and Kurohashi, S. (2017). "An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation." In Barzilay, R. and Kan, M.-Y. (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 385–391, Vancouver, Canada. Association for Computational Linguistics.

Dalvi, F., Durrani, N., Sajjad, H., and Vogel, S. (2018). "Incremental Decoding and Training Methods for Simultaneous Translation in Neural Machine Translation." In Walker, M., Ji, H., and Stent, A. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). "MuST-C: A Multilingual Speech Translation Corpus." In Burstein, J., Doran, C., and Solorio, T. (Eds.), *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)*, pp. 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Doi, K., Ko, Y., Makinae, M., Sudoh, K., and Nakamura, S. (2024). "Word Order in English-Japanese Simultaneous Interpretation: Analyses and Evaluation using Chunk-wise Monotonic Translation." In Salesky, E., Federico, M., and Carpuat, M. (Eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 254–264, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Doi, K., Sudoh, K., and Nakamura, S. (2021). "Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data." In Federico, M., Waibel, A., Costa-jussà, M. R., Niehues, J., Stuker, S., and Salesky, E. (Eds.), *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pp. 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

Fügen, C., Waibel, A., and Kolss, M. (2007). "Simultaneous Translation of Lectures and Speeches." *Machine translation*, **21**, pp. 209–252.

Fukuda, R., Doi, K., Sudoh, K., and Nakamura, S. (2024). "Creation of Evaluation Data for Monotonic Translation toward the Realization of Simultaneous English-Japanese Machine Translation Faithful to the Source Speech." In *Proceedings of the 259th Meeting of Special Interest Group of Natural Language Processing (IPSJ-SIGNL), 2024-NL-259(14)*, pp. 1–6. (in Japanese).

Fukuda, R., Nishikawa, Y., Kano, Y., Ko, Y., Yanagita, T., Doi, K., Makinae, M., Sakti, S., Sudoh, K., and Nakamura, S. (2023). "NAIST Simultaneous Speech-to-speech Translation System for IWSLT 2023." In Salesky, E., Federico, M., and Carpuat, M. (Eds.), *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pp. 330–340, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Gu, J., Neubig, G., Cho, K., and Li, V. O. (2017). "Learning to Translate in Real-time with Neural Machine Translation." In Lapata, M., Blunsom, P., and Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1053–1062, Valencia, Spain. Association for Computational Linguistics.

He, H., Boyd-Graber, J., and Daumé III, H. (2016). "Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation." In Knight, K., Nenkova, A., and Rambow, O. (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 971–976, San Diego, California. Association for Computational Linguistics.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." *IEEE ACM Transactions Audio Speech Language Processing*, **29**, pp. 3451–3460.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (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)*, pp. 8229–8233.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the Association for Computational Linguistics*, **5**, pp. 339–351.

Kano, Y., Sudoh, K., and Nakamura, S. (2022). "Simultaneous Neural Machine Translation with Prefix Alignment." In Salesky, E., Federico, M., and Costa-jussà, M. (Eds.), *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 22–31, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Kano, Y., Sudoh, K., and Nakamura, S. (2024). "Average Token Delay: A Duration-aware Latency Metric for Simultaneous Translation." *Journal of Natural Language Processing*, **31** (3), pp. 1049–1075.

Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., and Dupoux, E. (2020). "Data Augmenting Contrastive Learning of Speech Representations in the Time Domain." *arXiv preprint arXiv:2007.00991.*

Ko, Y., Fukuda, R., Nishikawa, Y., Kano, Y., Sudoh, K., and Nakamura, S. (2023). "Tagged End-to-End Simultaneous Speech Translation Training Using Simultaneous Interpretation Data." In Salesky, E., Federico, M., and Carpuat, M. (Eds.), *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pp. 363–375, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Kobus, C., Crego, J., and Senellart, J. (2017). "Domain Control for Neural Machine Translation." In Mitkov, R. and Angelova, G. (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 372–378, Varna, Bulgaria. INCOMA Ltd.

Kudo, T. and Richardson, J. (2018). "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing." In Blanco, E. and Lu, W. (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium. Association for Computational Linguistics.

Liu, D., Spanakis, G., and Niehues, J. (2020a). "Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection." In *Proceedings Interspeech 2020*, pp. 3620–3624.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer,

L. (2020b). “Multilingual Denoising Pre-training for Neural Machine Translation.” *Transactions of the Association for Computational Linguistics*, **8**, pp. 726–742.

Luong, M.-T. and Manning, C. (2015). “Stanford Neural Machine Translation Systems for Spoken Language Domains.” In Federico, M., Stüker, S., and Niehues, J. (Eds.), *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pp. 76–79, Da Nang, Vietnam.

Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H., and Wang, H. (2019). “STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework.” In Korhonen, A., Traum, D., and Màrquez, L. (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3025–3036, Florence, Italy. Association for Computational Linguistics.

Ma, X., Dousti, M. J., Wang, C., Gu, J., and Pino, J. (2020a). “SIMULEVAL: An Evaluation Toolkit for Simultaneous Translation.” In Liu, Q. and Schlangen, D. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 144–150, Online. Association for Computational Linguistics.

Ma, X., Pino, J., and Koehn, P. (2020b). “SimulMT to SimulST: Adapting Simultaneous Text Translation to End-to-End Simultaneous Speech Translation.” In Wong, K.-F., Knight, K., and Wu, H. (Eds.), *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*, pp. 582–587, Suzhou, China. Association for Computational Linguistics.

Matsushita, K., Yamada, M., and Ishizuka, H. (2020). “An overview of the Japan National Press Club (JNPC) Interpreting Corpus.” *Invitation to Interpreting and Translation Studies*, **22**, pp. 87–94.

Mizuno, A. (2017). “Simultaneous Interpreting and Cognitive Constraints.” *Journal of College of Literature, Aoyama Gakuin University,*, **58**, pp. 1–28.

Nishikawa, Y. and Nakamura, S. (2023). “Inter-connection: Effective Connection between Pre-trained Encoder and Decoder for Speech Translation.” In *Proceedings Interspeech 2023*, pp. 2193–2197.

Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (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)*, pp. 551–556.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019).

"fairseq: A Fast, Extensible Toolkit for Sequence Modeling." In Ammar, W., Louis, A., and Mostafazadeh, N. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: A Method for Automatic Evaluation of Machine Translation." In Isabelle, P., Charniak, E., and Lin, D. (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Polák, P., Pham, N.-Q., Nguyen, T. N., Liu, D., Mullov, C., Niehues, J., Bojar, O., and Waibel, A. (2022). "CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022." In Salesky, E., Federico, M., and Costa-jussà, M. (Eds.), *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Post, M. (2018). "A Call for Clarity in Reporting BLEU Scores." In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (Eds.), *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rao, K., Sak, H., and Prabhavalkar, R. (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)*, pp. 193–199. IEEE.

Rei, R., Farinha, A. C., Zerva, C., van Stigt, D., Stewart, C., Ramos, P., Glushkova, T., Martins, A. F. T., and Lavie, A. (2021). "Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task." In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C. (Eds.), *Proceedings of the 6th Conference on Machine Translation*, pp. 1030–1040, Online. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). "COMET: A Neural Framework for MT Evaluation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702.

Ren, Y., Liu, J., Tan, X., Zhang, C., Qin, T., Zhao, Z., and Liu, T.-Y. (2020). "SimulSpeech: End-to-End Simultaneous Speech to Text Translation." In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pp. 3787–3796, Online. Association for Computational Linguistics.

Rousseau, A., Deléglise, P., and Estève, Y. (2012). "TED-LIUM: An Automatic Speech Recognition Dedicated Corpus." In *International Conference on Language Resources and Evaluation*, pp. 125–129.

Sellam, T., Das, D., and Parikh, A. (2020). "BLEURT: Learning Robust Metrics for Text Generation." In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). "Controlling Politeness in Neural Machine Translation via Side Constraints." In Knight, K., Nenkova, A., and Rambow, O. (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 35–40, San Diego, California. Association for Computational Linguistics.

Shimizu, H., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2013). "Constructing a Speech Translation System using Simultaneous Interpretation Data." In Zhang, J. Y. (Ed.), *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). "Multilingual Translation from Denoising Pre-training." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3450–3466.

Toyama, H., Matsubara, S., Ryu, K., Kawaguchi, N., and Inagaki, Y. (2004). "CIAIR Simultaneous Interpretation Corpus." In *Proceedings of Oriental COCOSDA*.

Tsiamas, I., Gállego, G. I., Escolano, C., Fonollosa, J., and Costa-jussà, M. R. (2022). "Pretrained Speech Encoders and Efficient Fine-tuning Methods for Speech Translation: UPC at IWSLT 2022." In Salesky, E., Federico, M., and Costa-jussà, M. (Eds.), *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Wang, C., Pino, J., Wu, A., and Gu, J. (2020). "CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus." In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4197–4203, Marseille, France. European Language Resources Association.

Zhang, J. and Zong, C. (2016). "Exploiting Source-side Monolingual Data in Neural Machine

Translation." In Su, J., Duh, K., and Carreras, X. (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545, Austin, Texas. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (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.

Zhao, J., Sudoh, K., Nakamura, S., Ko, Y., Doi, K., and Fukuda, R. (2024). "NAIST-SIC-Aligned: An Aligned English-Japanese Simultaneous Interpretation Corpus." In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12046–12052, Torino, Italia. ELRA and ICCL.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. (2020). "Rethinking Pre-training and Self-training." *Advances in Neural Information Processing Systems*, **33**, pp. 3833–3845.

Yuka Ko: She is now a Ph.D. student in the Human-AI Interaction Laboratory, Nara Institute of Science and Technology. She received her bachelor's degree in engineering from the Osaka Prefecture University in 2020 and her master's degree in 2022 from the Nara Institute of Science and Technology. She is a recipient of the JSPS Research Fellowship for Young Scientist DC2. Her research interests include speech translation, simultaneous translation, and spoken language processing.

Ryo Fukuda: He is a researcher at the Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories. He received his bachelor's degree in engineering from the Toyohashi University of Technology in 2019 and his master's degree and doctor's degree in engineering from the Nara Institute of Science and Technology in 2021 and 2023.

Yuta Nishikawa: He received his bachelor's degree in engineering from the National Institute of Technology, Nara College in 2022 and his master's degree in engineering from the Nara Institute of Science and Technology in 2024. Since 2024, he has been the CEO of CueBeX Inc.

Yasumasa Kano: He received his bachelor's degree in economics from Yokohama National University in 2019, and his master's and Ph.D. degrees

in engineering in 2022 and 2024, respectively, from Nara Institute of Science and Technology. He was the CTO of Wonder Palette Inc. in 2024. Since 2021, he has been the CEO of Tleez Inc. He is a member of ANLP.

Katsuhito Sudoh: He is a professor at Nara Women's University. He received his bachelor's degree in engineering in 2000, and his master's and Ph.D. degrees in informatics in 2002 and 2015, respectively, from Kyoto University. He worked with the NTT Communication Science Laboratories from 2002 to 2017 and the Nara Institute of Science and Technology from 2017 to 2024. He works on machine translation and natural language processing. He is a member of ACL, ISCA, ANLP, IPSJ, ASJ, and JSAI.

Sakriani Sakti: She is the Head of the Human-AI Interaction (HAI) Research Laboratory and a Full Professor at NAIST, Japan. She also holds adjunct positions at JAIST and the University of Indonesia and is a Visiting Research Scientist at RIKEN AIP. She earned her B.E. from the Bandung Institute of Technology, Indonesia (1999), and her M.Sc. and Ph.D. from the University of Ulm, Germany (2002, 2008). From 2003 to 2011, she was a researcher at ATR and NICT, Japan, before joining NAIST, where she advanced from Assistant Professor to her current role. She has been actively involved in international collaborations, including the Asian Pacific Telecommunity Project, A-STAR, and U-STAR. She has also served as an IEEE SLTC Committee Member and an Associate Editor for IEEE/ACM TASLP. Currently, she chairs ELRA/ISCA SIGUL and serves as the Convener of O-COCOSDA. Her research focuses on multilingual and multimodal spoken language processing, deep learning, and cognitive communication.

Satoshi Nakamura: He is a Professor Emeritus at the Nara Institute of Science and Technology NAIST, a full professor at the Chinese University of Hong Kong, Shenzhen CUHKSZ, and an honorary professor at the Karlsruhe Institute of Technology, Germany. He received his B.S. degree from the Kyoto Institute of Technology in 1981 and a Ph.D. from the Kyoto University in 1992. He was the Director of the ATR SLC laboratory from 2000 to 2008 and the Director General of Keihanna Research Laboratories, NICT. He was a professor and Director of the AHC Laboratory at NAIST from 2011 to 2023. He is currently a full professor at CUHKSZ. He was an elected board member of the ISCA in the period of June 2011–2019, an IEEE Signal Processing Magazine

Editorial Board member in the period 2012–2015, and an IEEE SPS Speech and Language Technical Committee Member in the period 2013–2015. He is an ATR Fellow, IPSJ Fellow, IEEE Fellow, and ISCA Fellow.

(Received August 6, 2024)
(Revised December 1, 2024)
(Accepted January 15, 2025)