End-to-end Speech Translation with Spoken-to-Written Style Conversion

Anonymous ACL submission

Abstract

End-to-end speech translation (ST), which translates speech in source language directly into text in target language by a single model, has attracted a great deal of attention in recent years. Compared to the cascade ST, it has the advantages of easier deployment, better efficiency, and less error propagation. Meanwhile, spoken-to-written style conversion has been proved to be able to improve cascaded ST by reducing the gap between the language style of speech transcription and bilingual corpora used for machine translation training. Therefore, it is desirable to integrate the conversion into end-to-end ST. In this paper, we propose a joint task of speech-to-written-style-text conversion and end-to-end ST, as well as an interactive-attention-based multi-decoder model for the joint task to improve end-to-end ST. Experiments on a Japanese-English lecture ST dataset and CoVoST 2 Native Japanese show that our models outperform a strong baseline on Japanese-English ST.

1 Introduction

Speech-to-text translation (ST) is the task of translating a speech in source language into a text in target language. Traditionally, it is performed with a cascade approach (Stentiford and Steer, 1988; Waibel et al., 1991), dividing the task into 2 steps of automatic speech recognition (ASR) and text machine translation (MT). On the contrary, end-to-end ST (Berard et al., 2016) directly generates translations from speech without an intermediate step. Compared to cascade ST, it has the advantages of easier deployment, better efficiency, and less error propagation.

Spoken-to-written style conversion refers to converting text in spoken language into text in written language with identical semantic meaning. It is shown that spoken-to-written style conversion improves the accuracy of cascaded speech translation (ST) as it reduces the gap between the spoken language in the transcription generated by ASR systems and the written language in the bilingual corpora used for training MT systems (Nakao et al., 2021). See Table 1 for such an example.

Spoken-style Transcription	どっちが主になるかっていうの、ちょっと分からん。
Written-style Transcription	どちらが主になるかというのは、よく分からない。
English Translation	I'm not sure which one is the main thing.

Table 1: An example on spoken-to-written style conversion (difference is underlined). Compared to the original spoken-style Japanese transcription which is in spoken language, the written-style transcription is closer to the language style in bilingual corpora for MT, thus it is more likely to be machine-translated properly.

However, because of the limitation of cascaded ST, it is desirable to integrate spoken-to-written style conversion into the end-to-end ST. Due to the lack of large-scale ST corpora, existing well-performed end-to-end ST methods require the assist from ASR and MT by performing pre-training or multi-task learning. Therefore, it is reasonable that spoken-to-written style conversion can improve end-to-end ST as well because the gap between the language style of speech transcription and data for MT training still exits in end-to-end ST.

The contributions of this paper are two-fold:

- We propose an interactive-attention-based multi-decoder model that integrates spoken-to-written style conversion into end-to-end ST.
- We construct a large-scale lecture domain ST dataset on the language pair of Japanese-English and verify the effectiveness of our model on the lecture ST dataset together with the CoVoST 2 Native Japanese dataset.

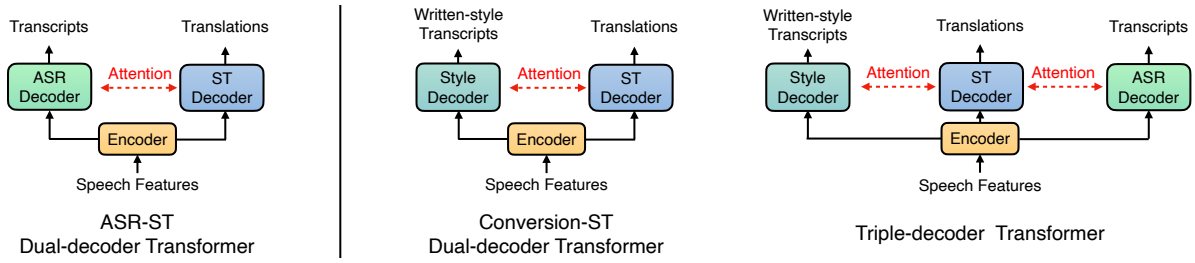

Figure 1: General structures of baseline (left) and proposed models (right two).

2 Related Work

Because the performance of simple end-to-end ST is generally limited, several works proposed using multi-task learning approach (Caruana, 1997) to improve the translation quality of end-to-end ST. Weiss et al. (2017) first investigated the subject and found that a multi-task model of ASR and ST performs better than that of MT and ST. In his proposal, only the encoder is shared and decoders for different tasks cannot utilize information from each. A two-stage model was proposed to alleviate the problem. (Kano et al., 2017; Anastasopoulos and Chiang, 2018; Sperber et al., 2019) It first performs ASR and then passes the decoder states as input to the ST decoder. However, it has the problem of limited efficiency of training and inference process.

Liu et al. (2020) proposed an interactive attention mechanism that enables ASR and ST to be performed synchronously. The ASR and ST decoders can exchange information with each other during the decoding process. Both decoders do not only rely on their previous outputs but also on the outputs produced by the other decoder.

Inspired by Liu et al. (2020)'s work, Le et al. (2020) presented dual-decoder Transformer. It has a more general framework with different variants and combinations of settings for the interactive attention mechanism. Our models are inspired by their work. We further improve the performance of their model by integrating the spoken-to-written conversion.

3 Method

Inspired by Le et al. (2020)'s work, we propose two model architectures: *conversion-ST dual-decoder Transformer* and *triple-decoder Transformer*. Correspondingly, the original dual-decoder Transformer is referred to as *ASR-ST dual-decoder Transformer*, which we use as a baseline.

Figure 1 shows the general structures of the baseline model and two proposed models. All the models are based on Transformer (Vaswani et al., 2017) and consist of one encoder and multiple sub-decoders. Each sub-decoder is specialized in producing outputs for a specific task.

3.1 Baseline

ASR-ST dual-decoder Transformer performs a joint task of ASR and ST, taking a sequence of speech features $\boldsymbol{s}$ as input and outputting a transcription $\boldsymbol{x}$ and a translation $\boldsymbol{y}$. The output distributions can be written as

$$
\begin{aligned}
D_{asr-st} &= p(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{s}) \\
&= \prod_{t=0}^{max(T_x, T_y)} p(x_t, y_t | \boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, \boldsymbol{s})
\end{aligned}
$$

The training objective is a weighted sum of cross-entropy losses for both tasks:

$$L_{asr-st} = \alpha L_{asr} + (1 - \alpha) L_{st}$$

α is set to 0.3 in all experiments.

3.2 Proposed Models

Conversion-ST dual-decoder Transformer performs a joint task of speech-to-written-style-text conversion and ST, taking a sequence of speech features $\boldsymbol{s}$ as input and outputting a written-style transcription $\boldsymbol{z}$ and a translation $\boldsymbol{y}$. Triple-decoder Transformer performs a joint task of ASR, speech-to-written-style-text conversion, and ST, taking a sequence of speech features $\boldsymbol{s}$ as input and outputting a transcription $\boldsymbol{x}$, a written-style transcription $\boldsymbol{z}$, and a translation $\boldsymbol{y}$. written-style transcription. Written-style transcription refers to the sentence that has the same semantic meaning as the transcription but is in written language instead of spoken language.

Similar to the ASR-ST dual-decoder, proposed models jointly predict their corresponding outputs in an autoregressive fashion. The output distributions can be written as

$$D_{conv-st} = p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{s}) = \prod_{t=0}^{max(T_y,T_z)} p(y_t, z_t|\boldsymbol{y}_{<t}, \boldsymbol{z}_{<t}, \boldsymbol{s})$$

$$D_{tri} = p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{s}) = \prod_{t=0}^{max(T_x,T_y,T_z)} p(x_t, y_t, z_t|\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, \boldsymbol{z}_{<t}, \boldsymbol{s})$$

The training objective for proposed models is a weighted sum of cross-entropy losses for the tasks that they perform:

$$L_{conv-st} = \alpha L_{conv} + (1-\alpha) L_{st}$$

$$L_{triple} = \alpha_1 L_{asr} + \alpha_2 L_{conv} + (1-\alpha_1-\alpha_2) L_{st}$$

In all experiments, α is set to 0.3, α_1 is set to 0.1, and α_2 is set to 0.2.

The purpose of proposing the triple-decoder and preserving the task of ASR is to compensate for possible information loss in written-style transcriptions. It should be noted that it is difficult to adopt our method directly as existing datasets for ST do not contain written-style transcriptions. It is more practical to synthesize written-style transcriptions from transcriptions with a pre-trained spoken-to-written style conversion model and use them as the ground truth for the task of speech-to-written-style-text conversion, which is the case in our experiments. In this case, the possible information loss is harder to ignore because the quality of the spoken-to-written style conversion is not fully guaranteed.

3.3 Interactive Attention Mechanism

Different decoders can exchange information with each other with the interactive attention mechanism, which refers to replacing attention sub-layers in the standard Transformer decoder with interactive attention sub-layers (Liu et al., 2020). In our models, the replaced sub-layers are the encoder-decoder attention sub-layers.

An interactive attention sub-layer consists of a main attention sub-layer and one or two cross attention sub-layers. It is only possible to have two cross attention sub-layers for the triple-decoder. The main attention sub-layer is the same as the replaced attention sub-layer. The cross attention sub-layers receives query $\boldsymbol{Q}$ from the same decoder A and receives key $\boldsymbol{K}$ and value $\boldsymbol{V}$ from another decoder B. $\boldsymbol{K}$ and $\boldsymbol{V}$ is from the same layer in decoder B, as in parallel dual-decoder. (Le et al., 2020) The final output is obtained by merging the output of the main attention sub-layer $\boldsymbol{H}_{main}$ with the output of the cross attention sub-layer $\boldsymbol{H}_{cross}$. We adopt a linear interpolation as the merging function, therefore the output representations of the interactive attention sub-layers with one and two cross attention sub-layers are

$$\boldsymbol{H}_{dual} = \boldsymbol{H}_{main} + \lambda \boldsymbol{H}_{cross}$$

$$\boldsymbol{H}_{triple} = \boldsymbol{H}_{main} + \lambda_1 \boldsymbol{H}_{cross1} + \lambda_2 \boldsymbol{H}_{cross2}$$

where $\boldsymbol{H}_{cross1}$ and $\boldsymbol{H}_{cross2}$ are outputs of two cross attention sub-layers that receives $\boldsymbol{K}$ and $\boldsymbol{V}$ from two different decoders. λ, λ_1, and λ_2 are all learnable parameters.

4 Experiments

4.1 Datasets

4.1.1 Datasets for ST

We constructed a dataset (Lecture ST) for our experiments with the data collected from lectures delivered at Kyoto University in 2019 and 2020. The raw data include full audio of the lectures, timestamps, Japanese transcriptions, and English translations from 15 courses. Timestamps mark the beginning and end of each transcription and each translation corresponds to multiple transcriptions. We combined timestamps and transcriptions to match the translations.

Some of the raw translations include multiple sentences. To segment them into sentence level, we first aligned the Japanese transcriptions to each English sentence based on the cosine similarity of their LASER[1] embeddings with the sentence alignment algorithm using dynamic programming proposed by Song et al. (2020). We then generated the new timestamps to align the text pairs with the audio with CTC segmentation (Kürzinger et al., 2020). Possible misalignment were manually checked and fixed in validation and test sets.

For ST experiments, we also used CoVoST 2 Native Japanese,[2] which is a rerecorded version of the Japanese section of CoVoST 2 (Wang et al., 2021) dataset all spoken by Japanese native speakers.

4.1.2 Datasets for ASR and MT Pre-training

We also constructed a lecture dataset for ASR pre-training (Lecture ASR) as we have data without

[1] https://github.com/facebookresearch/LASER
[2] https://github.com/ku-nlp/covost2NativeJa

Task	Dataset	Train	Valid	Test
ST	Lecture ST	71k	1,686	2,135
	CoVoST 2 NJ	1,119	635	684
ASR	Lecture ASR	170k	7,983	8,188
	CSJ	878K	8,644	8,622
MT	ASPEC-JE	2M	1,790	1,812

Table 2: The number of utterances in ST/ASR datasets and the number of sentences in the MT dataset.

Model	Lecture ST	CoVoST 2 NJ
ASR-ST Dual	25.08	2.37
Conv-ST Dual	25.81†	2.41
Triple	**26.02**†	**2.64**

Table 3: BLEU-4 scores on the two ST datasets. "†" indicates that the result is significantly better than "ASR-ST Dual" at $p < 0.05$).

translation from 6 extra courses. For this dataset, we used the raw timestamps and transcriptions from a total of 21 courses. We also used the 7th version of CSJ[3] for ASR pre-training. For MT pre-training, we used the Japanese-English section of ASPEC (ASPEC-JE) (Nakazawa et al., 2016). Table 2 shows the details of all the datasets that we used.

4.2 Spoken-to-Written Style Conversion

For the task of speech-to-written-style-text conversion, we need the written-style transcriptions corresponding to utterances in the datasets for ST. To obtain them, we trained a spoken-to-written style conversion model based on LaserTagger (Malmi et al., 2019) with the same data and settings as Nakao et al. (2021). The model gets a SARI (Xu et al., 2016) of 80.6 on the test set. We then used the model to synthesize written-style transcriptions from transcriptions in KU Lecture ST and CoVoST 2 Native Japanese.

4.3 Preprocessing and Model Settings

English translations were normalized and tokenized using the Moses tokenizer (Koehn et al., 2007). Japanese transcriptions and written transcriptions were tokenized using JUMAN++ (Morita et al., 2015; Tolmachev et al., 2018) and the punctuation was stripped. Japanese and English tokens were further split into subwords using the BPE method (Sennrich et al., 2016) with a joint vocabulary of 16k subwords.

[3]https://ccd.ninjal.ac.jp/csj/en/

Our implementation was based on the ESPnet-ST toolkit (Inaguma et al., 2020). For all the models, we used the same architecture with a 12-layer encoder and 8-layer decoders. For the triple-decoder, we only activated the cross attention sub-layers between the ST decoder and the other two decoders, as demonstrated in Figure 1. Encoders were initialized with an ASR model pre-trained on CSJ. For experiments on KU Lecture ST, the ASR model was further fine-tuned on KU Lecture ASR. ST decoders were initialized with an MT model pre-trained on ASPEC-JE. Other settings can be found in Appendix A.

4.4 Results

We report case-insensitive tokenized BLEU (Papineni et al., 2002) on two ST datasets. Significance tests were conducted using the bootstrap re-sampling method proposed by (Koehn, 2004). The results are shown in Table 3. Both proposed models (Conv-ST Dual and Triple) outperform the baseline (ASR-ST Dual) with triple-decoder making more improvement. We show more results for ablation study of interactive attention in Table 4 in Appendix B.

We analyzed the translation results with TER (Snover et al., 2006). The detailed statistics are shown in Table 5 in Appendix C. Compared to ASR-ST dual-decoder, conversion-ST dual-decoder tends to generate translations with fewer insertion and substitution errors as well as more deletion errors. It is intuitive because written-style transcriptions tend to be more concise but some of the information may be lost during the conversion. Triple-decoder can achieve a better trade-off with the assist of ASR and generate translations with relatively few errors in all categories.

5 Conclusion

In this paper, we proposed a joint task of speech-to-written-style-text conversion and end-to-end ST, as well as an interactive-attention-based multi-decoder architecture to perform the joint task. Compared to training on the joint task of ASR and ST, our method reduced the gap between the language styles of speech transcription and bilingual corpora used for MT pre-training. Experiments on Japanese-English ST datasets illustrated the effectiveness of our method. We plan to conduct experiments on other language pairs in the future.

References

Antonios Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1279–1283. ISCA.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *ArXiv*, abs/1612.01744.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 302–311. Association for Computational Linguistics.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2630–2634. ISCA.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer - 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7-9, 2020, Proceedings*, volume 12335 of *Lecture Notes in Computer Science*, pages 267–278. Springer.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3520–3533. International Committee on Computational Linguistics.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8417–8424. AAAI Press.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5053–5064. Association for Computational Linguistics.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Ryota Nakao, Chenhui Chu, and Sadao Kurohashi. 2021. Spoken-written japanese conversion for japanese-english university-lecture translation. *Journal of Natural Language Processing*, 28:1034–1052.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

Haiyue Song, Raj Dabre, Atsushi Fujita, and Sadao Kurohashi. 2020. Coursera corpus mining and multistage fine-tuning for improving lectures translation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3640–3649. European Language Resources Association.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Trans. Assoc. Comput. Linguistics*, 7:313–325.

Fred W.M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *1991 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '91, Toronto, Ontario, Canada, May 14-17, 1991*, pages 793–796. IEEE Computer Society.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Trans. Assoc. Comput. Linguistics*, 4:401–415.

A Experiment Details

We used the settings for speech features in the original dual-decoder Transformer (Le et al., 2020). The only difference was that we removed utterances having more than $6,000$ frames instead of $3,000$ because speakers can take long pauses within one sentence when delivering a lecture.

For the experiments on CoVoST 2 Native Japanese, the vocabulary was built with ASPEC-JE as building with CoVoST 2 Native Japanese led to a vocabulary that was too small for MT pre-training.

We used a Transformer with a 12-layer encoder and a 6-layer decoder for ASR pre-training and Transformer with a 6-layer encoder and an 8-layer decoder for MT training.

We used the Adam optimizer (Kingma and Ba, 2015) and Noam learning rate schedule (Vaswani et al., 2017) with 10k warm-up steps and a maximum learning rate of $1.5e-3$. We used a batch size of 24 sentences per GPU. All models were trained on a single machine with 8 Geforce GTX 1080 Ti GPUs. The models were trained for 75 epochs for experiments on KU Lecture ST and 150 epochs for experiments on CoVost 2 Japanese ST. We kept model checkpoints after each epoch and averaged the 5 best models on the validation set based on BLEU and used it for testing. For decoding, the beam size was set to 5 for the task of ST, 1 for the other tasks.

B Ablation Study

For ablation study, we conducted experiments on asymmetric models, which refers to only allowing the ST decoder to attend to other decoders, but not vise versa. We also conducted experiments in which we activate all of the cross attention sub-layers in 6 directions including the ones between the ASR decoder and the style decoder for the triple-decoder. The results are shown in Table 4. Changing to an asymmetric setting generally improves BLEU for the baseline (ASR-ST Dual asym) but not for proposed models (Conv-ST Dual asym, Triple asym), which shows that our proposed models are more dependent on bi-directional cross attention than the baseline. The results of triple-decoder with cross attention in 6 directions (Triple 6ca) are worse than that with cross attention in 4 directions, which shows that the cross attention between ASR decoder and style decoder is not necessary.

Model	Lecture ST	CoVoST 2 NJ
ASR-ST Dual	25.08	2.37
ASR-ST Dual asym	25.63	2.33
Conv-ST Dual	25.81	2.41
Conv-ST Dual asym	25.37	2.81
Triple	26.02	2.64
Triple asym	25.44	2.62
Triple 6ca	25.41	2.51

Table 4: BLEU using different settings of the interactive attention on the two ST datasets.

C TER

Table 5 shows the TER and detailed statistics of translations generated by the baseline and proposed models on 2 datasets.

Dataset	Model	Ins	Del	Sub	Shft	NumEr	NumWd	TER
Lecture ST	ASR-ST Dual	2,387	3,307	3,617	411	9,722	12,153	79.99
	Conv-ST Dual	2,340	3,456	3,510	400	9,706	12,153	79.86
	Triple	2,384	3,227	3,571	411	9,593	12,153	78.93
CoVoST 2 NJ	ASR-ST Dual	553	1,363	2,930	131	4,977	5,049	98.57
	Conv-ST Dual	507	1,404	2,891	134	4,936	5,049	97.76
	Triple	547	1,379	2,877	132	4,935	5,049	97.74

Table 5: TER and detailed statistics includes insertion (Ins), deletion (Del), substution (Sub), shift (Shft), number of errors (NumEr) and number of words (NumWd).