HW-TSC's Simultaneous Speech Translation System for IWSLT 2024

Shaojun Li, Zhiqiang Rao, Bin Wei, Yuanchang Luo, Zhanglin Wu ,
Zongyao Li, Hengchao Shang, Jiaxin Guo, Daimeng Wei, Hao Yang
Huawei Translation Service Center, Beijing, China
{lishaojun18, raozhiqiang, weibin29, luoyuanchang1, wuzhanglin2,
lizongyao, shanghengchao, guojiaxin1, weidaimeng, yanghao30}@huawei.com

Abstract

This paper outlines our submission for the IWSLT 2024 Simultaneous Speech-to-Text (SimulS2T) and Speech-to-Speech (SimulS2S) Translation competition. We have engaged in all four language directions and both the SimulS2T and SimulS2S tracks: English-German (EN-DE), English-Chinese (EN-ZH), English-Japanese (EN-JA), and Czech-English (CS-EN). For the S2T track, we have built upon our previous year's system and further honed the cascade system composed of ASR model and MT model. Concurrently, we have introduced an end-to-end system specifically for the CS-EN direction. This end-to-end (E2E) system primarily employs the pre-trained seamlessM4T model. In relation to the SimulS2S track, we have integrated a novel TTS model into our SimulS2T system. The final submission for the S2T directions of EN-DE, EN-ZH, and EN-JA has been refined over our championship system from last year. Building upon this foundation, the incorporation of the new TTS into our SimulS2S system has resulted in the ASR-BLEU surpassing last year's best score.

1 Introduction

This paper delineates the HW-TSC's contributions to the SimulS2T and SimulS2S Translation task at IWSLT 2024. Presently, research on SimulS2T translation from a systems architecture standpoint can be segregated into two categories: cascade and end-to-end. Cascade systems traditionally encompass a streaming Automatic Speech Recognition (ASR) module and a streaming text-to-text machine translation (MT) module, with an additional option of integrating correction modules. Despite the complexity of module integration, training each unit with ample data resources can yield significant results. On the other hand, an end-to-end approach for SimulS2T is feasible, where translations are directly procured from a unified model with speech input. It's worth mentioning, however, that bilingual speech translation datasets, indispensable for end-to-end models, remain scant.

Present efforts in simultaneous SimulS2T focus on the development of dedicated models customised for this task. This approach, nonetheless, comes with certain limitations, such as the need for an extra model, often accompanying a more complex training and inference process, augmented computational demands, and potential performance decrement when employed in an offline environment.

Our methodology for SimulS2T encompasses the use of a reliable offline ASR model and a robust offline MT model as the system's bedrock. We have adapted the onlinization approach of (Polák et al., 2022) and introduced an improved technique suitable for integration into the cascade system. On the official development set, our SimulS2T achieved a comparable level to the offline models under stringent latency constraints without any alterations to the original models. The disparity between offline and cascade has been further reduced compared to our last year's system. For the new CS-EN language pair, we submitted the end-to-end (E2E) system. We anticipate that future research will further enhance the E2E system's performance. Lastly, for SimulS2S, our system from the previous year had a low-performing TTS model. Hence, we updated the SimulS2S TTS model and integrated it with our latest SimulS2T system.

Our achievements is as follows:

- We further explored the upper limit of incremental decoding on our last year's champion SimulS2T system, and the BLEU value has been further improved compared to last year.
- We tried to extend our cascade SimulS2T method to the end-to-end system, and achieved the same effect with small losses between the offline and simultaneous system.

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 322–327
August 15-16, 2024

- Our method can be naturally extended to the SimulS2S system, and after SimulS2T reduced minor error propagation, SimulS2S achieved greater improvements.

2 Models

All models used by our system are offline models and do not use special streaming strategies. The following is an introduction to each model.

2.1 Offline ASR

In all our cascade system, Our system uses the U2 (Wu et al., 2021) framework as the ASR (Automatic Speech Recognition) module because it is flexible and supports streaming and non-streaming ASR. U2's key features include dynamic chunk training, CTC decoder, and autoregressive attention decoder. It's capable of conditional training with different chunk sizes and allows for multiple decoding strategies. We use "attention_rescoring" for re-scoring CTC generated texts.

2.2 Offline MT

The Machine Translation (MT) module of our system is the Transformer (Vaswani et al., 2017), a very common tool used in machine translation (Wei et al., 2021; Li et al., 2022). To improve this, we use multiple training strategies like multilingual translation (Johnson et al., 2017) for English to German, Chinese, and Japanese, forward translation (Wu et al., 2019) for generating synthetic data (Nguyen et al., 2020), and generation from an ASR model to reduce the domain gap.

2.3 Offline S2T

For CS-EN direction, We used the offline SeamlessM4T (Seamless Communication, 2023) as our end-to-end SimulS2T model. SeamlessM4T integrates a deep learning framework with a self-supervised speech representation learned from 1 million hours of open speech audio data using w2v-BERT 2.0. The speech to text model employs a audio encoder and text decoder. Open-sourced for community development, SeamlessM4T includes robust safety measures to mitigate harmful content and is designed to be adaptable for various applications, from international communication to content creation.

2.4 Offline TTS

The Text-to-Speech (TTS) module is vital for generating high-quality speech from translated text. We use the VITS (Kim et al., 2021) model for this, a state-of-the-art tool that can produce natural, fluent speech. The process is efficient, only needing the generated text to create the raw audio waveform. This makes the TTS module faster and improves the user experience.

3 Methods

3.1 Cascaded SimulS2T

For EN-DE, EN-JA, EN-ZH, we followed last year's model (Guo et al., 2023). Regarding the incremental decoding strategy, we added vad segmentation and chunk padding on the ASR side to achieve smaller delays, and added an ensemble strategy on the MT side to achieve better MT stability.

Onlinization Incremental Decoding is the main way to make an offline model into a real-time one. Translation tasks might need reordering or more information, which isn't clear until the source sentence ends. Offline models work best when they can process the whole sentence at once, but this can cause delays in real-time mode. A possible solution is to break the source sentence into smaller pieces and translate each piece separately. This lessens the processing time while keeping the translation quality. By using incremental decoding with these smaller pieces, we can speed up the translation process a lot, which is perfect for real-time situations.

In incremental inference, we break the input sentence into set-sized chunks and decode each chunk as it comes in. Once a chunk is chosen, its predictions are locked in and aren't changed anymore to avoid distractions from constantly changing guesses. The decoding of the next chunk depends on the locked-in predictions. In reality, decoding for new chunks can either continue from a saved decoder state or start after forced decoding with the locked-in tokens. In either situation, the source-target attention can cover all available chunks, not just the current one.

Prefix Vad Incremental decoding can pose challenges with longer sentences. As the sentence lengthens and the prefix extends, the decoding process tends to slow down, relying progressively on extensive contexts. Consequently, this requires waiting for more chunks to output translation results, which in turn affects our decoding delay. To mitigate this, we propose incorporating vad (Tong et al., 2014) detection and trimming excessively long prefixes once the input reaches a sufficient length. This strategy helps to minimize the streaming delay and reduce computational overhead for the model. Simultaneously, to ensure decoding quality, it's crucial to maintain adequate context. Therefore, when detecting vad, we consider the current vad position's distance from the sentence's start and end. Balancing this length setting with overall performance is a key aspect of our approach.

Chunk Padding During the ASR streaming decoding process, we observed instability with decoding the final few frames of the audio features. This instability presumably results from the model's insufficient edge handling during the convolution process. This issue consequently disrupts the beam search of streaming ASR, leading to inconsistent or sometimes erroneous outcomes. These errors are then carried over to the MT model, negatively impacting the streaming translation's stability. However, by appending blank padding to the end of each streaming chunk, we can notably enhance the decoding stability for the stream's last few words.

MT Ensemble In cascaded systems, the elimination of error propagation is often challenging. The erroneous ASR inputs that the MT system processes often lead to more significant errors. Moreover, our MT system has certain constraints due to its use of various strategies for domain adaptation and fine-tuning, resulting in an overfitting risk. To address these issues, we have incorporated MT models trained with diverse strategies into this year's system. By employing ensemble (Sagi and Rokach, 2018) methods, we aim to enhance the model's fault tolerance while simultaneously mitigating the risk of overfitting in the field.

3.2 E2E SimulS2T

For the newly introduced language direction this year, CS-EN, we utilized the pre-trained seamlessM4T as our end-to-end SimulS2T model. We attempted to fine-tune the seamlessM4T using the officially provided data. Concurrently, we implemented the aforementioned cascaded SimulS2T decoding strategy to seamlessM4T, aiming to attain a streaming effect.

3.3 Cascaded SimulS2S

In a cascaded speech-to-speech translation system, the TTS module plays a critical role in rendering high-quality speech output from translated text. To this end, we utilize the state-of-the-art VITS (Kim et al., 2021) model, which is pretrained on massive amounts of data and incorporates advanced techniques such as variational inference augmented with normalizing flows and adversarial training. This model has been shown to produce speech output that is more natural and fluent compared to traditional TTS models.

The inference process involves providing the VITS model with the generated text, after which the model generates the raw audio waveform. This process is highly efficient and requires no additional input from the user. By leveraging the VITS model, we are able to streamline the TTS module and deliver high-quality speech output in a fraction of the time traditionally required by other systems. This results in a more seamless and intuitive user experience, enabling our system to be used by a wider range of individuals and applications.

4 Experiments Setup

4.1 Dataset

We used four datasets to train the ASR (Automatic Speech Recognition) module: LibriSpeech V12, MuST-C V2, TEDLIUM V3, and CoVoST V2. Each dataset contains different types of data, like audio book recordings, TED talks, and open-domain content. LibriSpeech doesn't have punctuations in its texts, but MuST-C and CoVoST do.

For training the machine translation (MT) model, we collected all available parallel corpora that were similar to the MuST-C domain, then trained a multilingual MT baseline model. We also incrementally trained the model based on data from each language direction.

4.2 Model

ASR We used 80-dimensional Mel-Filter bank features from audio files to create the ASR training corpus, and Sentencepiece for ASR texts tokenization. The ASR model has different configurations for encoder layers, decoder layers, heads, hidden dimensions, and FFN. For training, we used a batch size of up to 40,000 frames per card and trained the model on 4 GPUs for 50 epochs. To improve accuracy, we augmented all audio inputs with spectral augmentation and normalized with utterance cepstral mean and variance normalization. We also apply prefix vad and chunk padding in the asr decoding metioned in Section 3.1.

System	Language Pair	BLEU	AL	AP	DAL
IWSLT23 Best	EN-DE	33.54	1.88	0.83	2.84
Our System	EN-DE	**34.24**	1.94	0.84	2.94
IWSLT23 Best	EN-JA	17.89	1.98	0.83	2.89
Our System	EN-JA	**17.94**	1.93	0.84	2.82
IWSLT23 Best	EN-ZH	27.23	1.98	0.83	2.89
Our System	EN-ZH	**27.63**	1.88	0.83	2.82
Our System	CS-EN	19.03	1.96	0.91	3.67

Table 1: Final systems results of SimulS2T on tsc-Common v2.0/dev

Model	Language Pair	ASR_BLEU	StartOffset	EndOffset	ATD
IWSLT23 Best	EN-DE	26.7	2.33	5.67	-
Our System	EN-DE	**27.09**	1.86	4.17	3.06
IWSLT23 Best	EN-JA	14.53	1.59	2.96	2.76
Our System	EN-JA	**15.55**	2.32	3.15	2.89
IWSLT23 Best	EN-ZH	20.19	1.77	2.98	2.93
Our System	EN-ZH	**22.92**	1.76	3.0	2.79
Our System	CS-EN	17.12	1.48	4.11	4.09

Table 2: Final systems results of SimulS2S on tsc-Common v2.0/dev

MT We used the Transformer deep model architecture for our MT model experiments. The configuration of this model includes encoder layers, decoder layers, heads, hidden dimensions, FFN, and pre_ln. The model was trained using 8 GPUs, with a batch size of 2048, a parameter update frequency of 32, and a learning rate of 5e-4. During inference, we used a beam size of 8 and set the length penalties to 1.0. We selected 2 MT models for ensemble mentioned in Section 3.1.

S2T We finetuned seamlessM4T-medium with the official data, BLEU improved by two points but did not exceed seamlessM4T-large-v2. Finally, we submitted the seamlessM4T-large-v2 model as our E2E model. We used the same decoding strategy as the cascade, using a beam_size of 5 and setting no_repeat_ngram_size.

TTS For EN-DE direction, we utilize the open-source Espnet (Watanabe et al., 2018) for inference. For EN-JA/ZH and CS-EN, we use the pretrained models in huggingface. The pretrained models are VITS (Kim et al., 2021) architecture, which adopts variational inference augmented with normalizing flows and an adversarial training process.

5 Results

5.1 SimulS2T

From Table 1, we can see that the our systems work well on various language pairs. And our systems even beat the best IWSLT23 systems of ourselves with methods mentioned in Section 3. EN-DE has improved the most. Since the gap between EN-DE offline and streaming is much larger than that of EN-JA and EN-ZH, we found that there is still a big gap between the MT results of cascaded streaming and ASR golden. In subsequent research, we may focus on this point.

5.2 SimulS2S

From Table 2, we observed that the improvement of S2S is greater than that of S2T. For EN-DE, most of the improvement is mainly due to our replacement of the TTS model, while for EN-JA and EN-ZH, thanks to the more stable SimulS2T, we spread to The TTS error is smaller, so the improvement of SimulS2S is more obvious than SimulS2T, which also illustrates the impact of error propagation on cascade system.

5.3 Ablation Study on Decoding Strategies

Decoding strategies	BLEU
Baseline	34.24
IWSLT23 Best	33.54
- Prefix Vad	33.91
- Chunk Padding	34.02
- Ensemble	33.86

Table 3: Ablation Study on Decoding Strategies

We separately studied the impact of today's newly introduced decoding strategies on translation quality: prefix vad, chunk padding, ensemble. It is evident from Table 3 that these decoding strategies can effectively improve the overall quality of the system.

6 Conclusion

In summary, this paper presents our efforts for the IWSLT 2024 Simultaneous Speech-to-Text and Speech-to-Speech Translation competition. We improved upon our previous system, achieved better translation accuracy and successfully integrated a novel Text-to-Speech model. Our system uses reliable offline models, and we managed to enhance the simulated conversation translation system's quality. Our experiments demonstrated that our system performs well across different language pairs. Future work will pay more focus on end-to-end systems.

References

Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Zongyao Li, Zhiqiang Rao, Minghan Wang, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Shaojun Li, Yuhao Xie, Lizhi Lei, and Hao Yang. 2023. The HW-TSC's simultaneous speech-to-text translation system for IWSLT 2023 evaluation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 376–382, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Shaojun Li, Yuanchang Luo, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Yuhao Xie, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC systems for WMT22 very low resource supervised MT task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1098–1103, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq R. Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Peter Polák, Ngoc-Quan Pham, Tuan-Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondrej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation, IWSLT@ACL 2022, Dublin, Ireland (in-person and online), May 26-27, 2022*, pages 277–285. Association for Computational Linguistics.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1249.

Yu-An Chung Mariano Coria Meglioli David Dale Ning Dong Mark Duppenthaler Paul-Ambroise Duquenne Brian Ellis Hady Elsahar Justin Haaheim John Hoffman Min-Jae Hwang Hirofumi Inaguma Christopher Klaiber Ilia Kulikov Pengwei Li Daniel Licht Jean Maillard Ruslan Mavlyutov Alice Rakotoarison Kaushik Ram Sadagopan Abinesh Ramakrishnan Tuan Tran Guillaume Wenzek Yilin Yang Ethan Ye Ivan Evtimov Pierre Fernandez Cynthia Gao Prangthip Hansanti Elahe Kalbassi Amanda Kallet Artyom Kozhevnikov Gabriel Mejia Robin San Roman Christophe Touret Corinne Wong Carleigh Wood Bokai Yu Pierre Andrews Can Balioglu Peng-Jen Chen Marta R. Costa-jussà Maha Elbayad Hongyu Gong Francisco Guzmán Kevin Heffernan Somya Jain Justine Kao Ann Lee Xutai Ma Alex Mourachko Benjamin Peloquin Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Anna Sun Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Mary Williamson Seamless Communication, Loïc Barrault. 2023.

Seamless: Multilingual expressive and streaming speech translation.

Sibo Tong, Nanxin Chen, Yanmin Qian, and Kai Yu. 2014. Evaluating vad for automatic speech recognition. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 2308–2314. IEEE.

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.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. *CoRR*, abs/1804.00015.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. Hw-tsc's participation in the WMT 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 225–231. Association for Computational Linguistics.

Di Wu, Binbin Zhang, Chao Yang, Zhendong Peng, Wenjing Xia, Xiaoyu Chen, and Xin Lei. 2021. U2++: unified two-pass bidirectional end-to-end model for speech recognition. *CoRR*, abs/2106.05642.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4205–4215. Association for Computational Linguistics.