The HW-TSC's Simultaneous Speech-to-Speech Translation system for IWSLT 2023 evaluation

Hengchao Shang, Zhiqiang Rao, Zongyao Li, Jiaxin GUO, Zhanglin Wu, Minghan Wang,

Daimeng Wei, Shaojun Li, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang

{shanghengchao, raozhiqiang, lizongyao, guojiaxin1, wuzhanglin2, wangminghan, weidaimeng, lishaojun18, yuzhengzhe, chenxiaoyu35, leilizhi, yanghao30}@huawei.com

Abstract

In this paper, we present our submission to the IWSLT 2023 (Agarwal et al., 2023) Simultaneous Speech-to-Speech Translation competition. Our participation involves three language directions: English-German, English-Chinese, and English-Japanese. Our solution is a cascaded incremental decoding system, consisting of an ASR model, an MT model, and a TTS model. By adopting the strategies used in the Speech-to-Text track, we have managed to generate a more confident target text for each audio segment input, which can guide the next MT incremental decoding process. Additionally, we have integrated the TTS model to seamlessly reproduce audio files from the translation hypothesis. To enhance the effectiveness of our experiment, we have utilized a range of methods to reduce error conditions in the TTS input text and improve the smoothness of the TTS output audio.

1 Introduction

This paper describes the HW-TSC's submission to the Simultaneous Speech-to-Speech Translation (SimulS2S) task at IWSLT 2023 (Agarwal et al., 2023).

Simultaneous speech-to-speech translation (SimulS2S) is currently being researched using Cascade systems. These systems typically involve a streaming Automatic Speech Recognition (ASR) module, a streaming Text-to-Text machine translation (MT) module, and an offline Text-to-Speech(TTS) module, with the option of incorporating additional correction modules. Although integrating these modules can be complex, training each module with sufficient data resources can prove to be worthwhile.

Our study adopts a comprehensive approach that utilizes several key components to build a strong system. We incorporate a formidable offline ASR model, a robust offline MT model, and a pre-trained TTS model as the foundation for our system. Moreover, we introduce a refined onlinization technique based on the approach developed by (Polák et al., 2022), which seamlessly integrates into the cascade system.

Offline TTS models often produce a blank sound at the end of a sentence. As a result, when generating audio results in the simultaneous interpreting mode, it can lead to blank tones between clips, causing the final audio to lack smoothness. To address this issue, we have developed several strategies aimed at mitigating this problem in our work.

2 Related Methods

2.1 ASR

In our cascade system, we have incorporated the U2 (Wu et al., 2021) as the ASR module. This framework has the flexibility to be implemented on standard Transformer or Conformer architectures and can perform both streaming and non-streaming ASR. One of the major advantages of U2 over other offline autoregressive ASR models is its ability to support streaming through dynamic chunk training and decoding with a CTC decoder on top of the encoder. Additionally, U2 includes a standard autoregressive attention decoder and can be jointly trained with the CTC decoder to improve training stability. The dynamic chunk training method involves applying a causal mask with varying chunk sizes at the self-attention layer within the encoder. This allows the hidden representation to condition on some look-ahead contexts within the chunk, similar to the self-attention of an autoregressive decoder.

U2 offers multiple decoding strategies. In this work, we use "attention_rescoring" decoding strategy, which is to use the attention decoder re-score CTC generated texts using prefix beam search in the event of multiple candidate proposals.

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 383–388

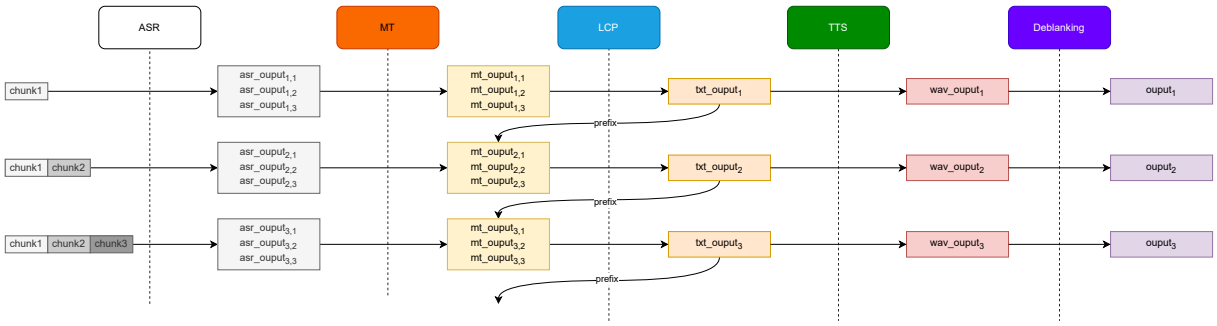

Figure 1: An overview of hw-tsc's s2s framework.

2.2 MT

Our cascade system includes the Transformer (Vaswani et al., 2017) as the MT module, which has become a prevalent method for machine translation (Wei et al., 2021; Guo et al., 2021) in recent years. The Transformer has achieved impressive results, even with a primitive architecture that requires minimal modification.

In this work, we use multiple training strategies to improve the offline MT model performance. First, we train a multilingual model for three directions En-De/ZH/JA. Multilingual Translation (Johnson et al., 2017) has proposed a simple solution to enhance translation performance for translating multiple languages using a single neural machine translation model with no need to alter the model architecture. Second, we use Forward translation (Wu et al., 2019) to generate synthetic data through beam search decoding. The we add the data to the original parallel corpora and re-train the MT model. Finally, we use the generation from a well-trained ASR model to replace source-side text in the training corpus data and fine-tune the MT model to reduce the domain gap.

2.3 TTS

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.

3 Framework

Figure 1 illustrates our framework.

3.1 Onlinization

The primary method for onlinizing an offline model and transforming it into a simul model is Incremental Decoding. Depending on the language pair, translation tasks may require reordering or additional information that is not apparent until the end of the source utterance. In offline settings, processing the entire utterance at once usually produces the highest-quality results, but this approach can result in significant latency in online mode. One possible solution to reduce latency is to divide the source utterance into smaller parts and translate each part separately. This approach helps to reduce the time required for processing while still maintaining translation quality. By using incremental decoding in conjunction with smaller processing units, we can significantly improve the speed and efficiency of the translation process, making it ideal for online settings where speed is of the essence.

To perform incremental inference, we divide the input utterance into chunks of a fixed size and decode each chunk as it arrives. Once a chunk has been selected, its predictions are then committed to and no longer modified to avoid visual distractions from constantly changing hypotheses. The decoding of the next chunk is dependent on the pre-

dictions that have been committed to. In practice, decoding for new chunks can proceed from a previously buffered decoder state or begin after forced decoding with the tokens that have been committed to. In either case, the source-target attention can span all available chunks, as opposed to only the current chunk.

3.2 Stable Hypothesis Detection

Our approach is based on prior research in (Polák et al., 2022), and we have implemented stable hypothesis detection to minimize the potential for errors resulting from incomplete input. In previous research, some methods focused on detecting stable hypotheses using strategies such as the Hold-n strategy proposed by (Liu et al., 2020), which identifies the best hypothesis in the beam and removes the last n tokens from it. Similarly, (Liu et al., 2020) introduced the LA-n strategy, which identifies the matching prefixes of two consecutive chunks. In addition, (Nguyen et al., 2021) developed the SP-n strategy, which identifies the longest common prefix among all items in the beam of a chunk.

However, these methods were designed for end-to-end systems that search for a shared prefix among the hypotheses generated from different chunk inputs. Our approach, on the other hand, operates within a cascaded system that processes the same chunk input. As such, we have adapted these strategies to better fit our context, resulting in a more effective approach for stable hypothesis detection. By using our approach, we are able to achieve higher accuracy and stability in our system, thereby improving its overall performance.

We can denote the MT and ASR generating functions as G and F respectively. Let $F^C_{i,n}$ represent the i output generated by the ASR function for a c-chunk input with a beam size of n. Then the final common prefix for the c-chunk input can be expressed as $prefix^c$, which is determined as follows:

$$prefix^c = LCP(G(F^c_{1,n}), ..., G(F^c_{n,n})) \quad (1)$$

where $LCP(\cdot)$ is longest common prefix of the arguments.

3.3 Deblanking

Our team conducted a manual evaluation of the audio output generated by TTS and identified two issues. The first scenario involved the TTS model producing unusual waveforms for previously unseen tokens. The second scenario involved TTS generating blank sounds to indicate pauses within the audio fragments. To address these issues, we implemented two strategies which we have collectively named Deblanking.

Unknown Filtering In the Chinese and Japanese language directions, we initially remove tokens that are not included in the vocabulary, such as infrequent punctuation marks and words. For Chinese in particular, we must convert Arabic numerals into textual numerals.

Context-Aware Pause Detection When analyzing the waveform generated by TTS, we evaluate whether or not the original text indicates a pause. If the text does not indicate a pause, we eliminate the final prolonged silence that produces the waveform. Additionally, to ensure speech coherence, we've reserved at least 160 frames of blank audio.

4 Experiments

4.1 Dataset

To train the ASR module, we utilized four datasets: LibriSpeech V12, MuST-C V2 (Gangi et al., 2019), TEDLIUM V3, and CoVoST V2. LibriSpeech consists of audio book recordings with case-insensitive text lacking punctuation. MuST-C, a multilingual dataset recorded from TED talks, was used solely for the English data in the ASR task. TEDLIUM is a large-scale speech recognition dataset containing TED talk audio recordings along with text transcriptions. CoVoST is also a multilingual speech translation dataset based on Common Voice, with open-domain content. Unlike LibriSpeech, both MuST-C and CoVoST have case-sensitive text and punctuation.

To train the MT model, we collected all available parallel corpora from the official websites and selected data that was similar to the MuST-C domain. We first trained a multilingual MT baseline model on all data from three language directions. Then, we incrementally trained the baseline model based on data from each language direction.

4.2 Model

ASR We extract 80-dimensional Mel-Filter bank features from audio files to create the ASR training corpus. For tokenization of ASR texts, we utilize Sentencepiece with a learned vocabulary of up to

Model	Language Pair	BLEU/Whisper_ASR_BLEU	StartOffset	EndOffset	ATD
Our S2T System	EN-DE	33.54			
	EN-JA	17.89			
	EN-ZH	27.23			
Our System	EN-DE	10.45	1.04	2.73	1.97
Our System	EN-JA	14.53	1.59	2.96	2.76
Our System	EN-ZH	20.19	1.77	2.98	2.93

Table 1: Final systems results

20,000 sub-tokens. The ASR model is configured as follows: $n_{encoder\ layers}$ = 12, $n_{decoder\ layers}$ = 8, n_{heads} = 8, d_{hidden} = 512, d_{FFN} = 2048. We implement all models using wenet (Zhang et al., 2022).

During the training of the ASR model, we set the batch size to a maximum of 40,000 frames per card. We use inverse square root for lr scheduling, with warm-up steps set to 10,000 and peak lr set at $5e-4$. Adam is utilized as the optimizer. The model is trained on 4 V100 GPUs for 50 epochs, and the parameters for the last 4 epochs are averaged. To improve accuracy, all audio inputs are augmented with spectral augmentation and normalized with utterance cepstral mean and variance normalization.

MT For our experiments using the MT model, we utilize the Transformer deep model architecture. The configuration of the MT model is as follows: $n_{encoder\ layers}$ = 25, $n_{decoder\ layers}$ = 6, n_{heads} = 16, d_{hidden} = 1024, d_{FFN} = 4096, $pre\_ln = True$.

We utilize the open-source Fairseq (Ott et al., 2019) for training, with the following main parameters: each model is trained using 8 GPUs, with a batch size of 2048, a parameter update frequency of 32, and a learning rate of $5e-4$. Additionally, a label smoothing value of 0.1 was used, with 4000 warmup steps and a dropout of 0.1. The Adam optimizer is also employed, with $\beta 1$ = 0.9 and $\beta 2$ = 0.98. During the inference phase, a beam size of 8 is used. The length penalties are set to 1.0.

TTS For EN-DE direction, we utilize the open-source Espnet (Watanabe et al., 2018) for inference. For EN-JA/ZH, 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.

4.3 Results

A detailed analysis of the results presented in Table 1 indicates that the TTS transcription results in Japanese have the smallest gap compared to the results obtained from the S2T system, with a difference of approximately 3 BLEU. However, in the German direction, the TTS system generates the worst results among all the evaluated systems. Further research is needed to understand the underlying reasons for this discrepancy and identify potential strategies to improve TTS performance in this language pair.

4.4 Ablation Study on Deblanking strategies

Language Pair	Training strategies	BLEU
EN-DE	Baseline	10.45
	- Context-aware wait	10.32
	- Unknown Filtering	10.27
EN-JA	Baseline	14.53
	- Context-aware wait	13.37
	- Unknown Filtering	13.08
EN-ZH	Baseline	20.19
	- Context-aware wait	18.64
	- Unknown Filtering	16.73

Table 2: Ablation Study on Deblanking strategies

The results presented in Table 2 provide strong evidence that our proposed strategies are effective in reducing the gap between offline and streaming TTS.

5 Conclusion

This paper details our involvement in the IWSLT 2023 simultaneous speech-to-speech translation evaluation. Our team presents an onlinization strategy that can be utilized by cascaded systems, which we have proven to be effective in three different language directions. Additionally, we introduce two strategies that address the disparity between

offline and streaming TTS. Our approach is both simple and efficient. Moving forward, we aim to delve further into end-to-end systems.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Association for Computational Linguistics.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 2012–2017. Association for Computational Linguistics.

Jiaxin Guo, Minghan Wang, Daimeng Wei, Hengchao Shang, Yuxia Wang, Zongyao Li, Zhengzhe Yu, Zhanglin Wu, Yimeng Chen, Chang Su, Min Zhang, Lizhi Lei, Shimin Tao, and Hao Yang. 2021. Self-distillation mixup training for non-autoregressive neural machine translation. CoRR, abs/2112.11640.

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.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, pages 3620–3624. ISCA.

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. 2021. Super-human performance in online low-latency recognition of conversational speech. In Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021, pages 1762–1766. ISCA.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. CoRR, abs/1904.01038.

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.

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.

Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu. 2022. Wenet 2.0: More productive end-to-end speech recognition toolkit. In Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022, pages 1661–1665. ISCA.