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

Jiaxin GUO, Daimeng Wei, Zhanglin Wu, Zongyao Li, Zhiqiang Rao, Minghan Wang,

Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Shaojun Li, Yuhao Xie, Lizhi Lei, Hao Yang

{guojiaxin1, weidaimeng, wuzhanglin2, lizongyao, raozhiqiang,
wangminghan, shanghengchao, chenxiaoyu35, yuzhengzhe,
lishaojun18, xieyuhao2, leilizhi, yanghao30}@huawei.com

Abstract

In this paper, we present our submission to the IWSLT 2023 (Agarwal et al., 2023) Simultaneous Speech-to-Text Translation competition. Our participation involves three language directions: English-German, English-Chinese, and English-Japanese. Our proposed solution is a cascaded incremental decoding system that comprises an ASR model and an MT model. The ASR model is based on the U2++ architecture and can handle both streaming and offline speech scenarios with ease. Meanwhile, the MT model adopts the Deep-Transformer architecture. To improve performance, we explore methods to generate a confident partial target text output that guides the next MT incremental decoding process. In our experiments, we demonstrate that our simultaneous strategies achieve low latency while maintaining a loss of no more than 2 BLEU points when compared to offline systems.

1 Introduction

This paper describes the HW-TSC‘s submission to the Simultaneous Speech-to-Text Translation (SimulS2T) task at IWSLT 2023 (Agarwal et al., 2023).

From a systems architecture perspective, current research on simultaneous speech-to-text translation (SimulS2T) can be categorized into two forms: cascade and end-to-end. Cascade systems typically consist of a streaming Automatic Speech Recognition (ASR) module and a streaming text-to-text machine translation (MT) module, with the possibility of incorporating additional correction modules. While integrating these modules can be complex, training each module with sufficient data resources can prove to be worthwhile. Alternatively, an end-to-end approach is also an option for SimulS2T, where translations can be directly generated from a unified model with speech inputs. However, it is important to note that bilingual speech translation datasets, which are necessary for end-to-end models, are still scarce resources.

The current efforts in simultaneous speech-to-text translation (SimulS2T) concentrate on developing dedicated models that are tailored to this specific task. However, this approach has certain drawbacks, such as the requirement of an additional model, which typically involves a more challenging training and inference process, as well as heightened computational demands and the possibility of decreased performance when utilized in an offline environment.

Our approach for this study involves utilizing a sturdy offline ASR model and a robust offline MT model as the foundation for our system. By modifying the onlinization approach of (Polák et al., 2022) and introducing an enhanced technique that can be seamlessly integrated into the cascade system, we are able to demonstrate that our simultaneous system can perform at the similar level as the offline models under strict latency restrictions without any adjustments to the original models. Furthermore, our system even surpasses previous higher latency IWSLT systems.

Our contribution is as follows:

- We have revised the approach of onlinization adopted by (Polák et al., 2022) and put forward an enhanced technique that can be easily integrated into the cascade system.
- Our findings show that the pre-training plus fine-tuning paradigm yields significant improvements in both ASR and MT.
- Our research highlights that enhancing the offline MT model has a direct positive impact on the online cascade system as well.

2 Related Work

Simultaneous speech-to-text translation can be achieved through either a cascaded system or an

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 376–382
July 13-14, 2023 ©2023 Association for Computational Linguistics

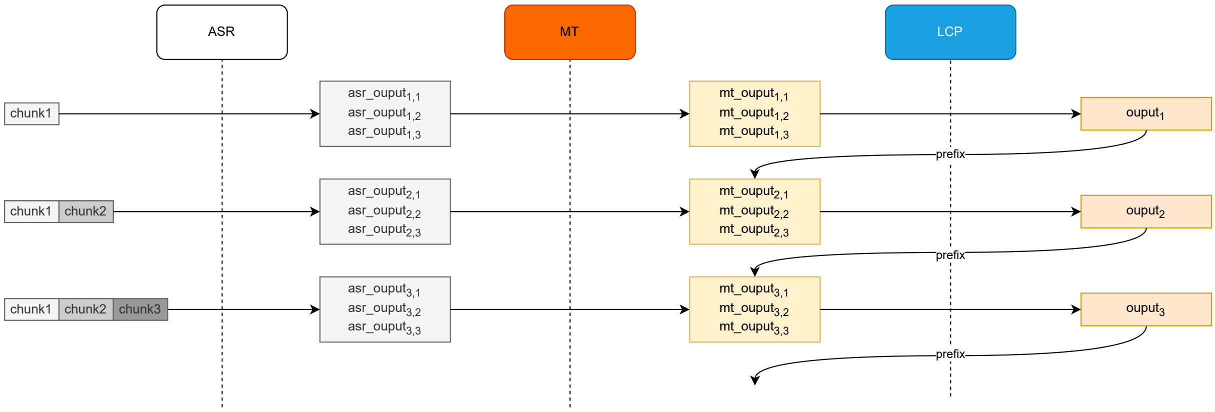

Figure 1: An overview of hw-tsc's s2t framework.

end-to-end model, both of which can be (hybrid) in nature. While cascaded systems currently offer the highest quality in offline speech translation, end-to-end speech translation provides a better trade-off between quality and latency (Guo et al., 2022; Wang et al., 2022a,b).

End-to-end speech translation systems incorporate various techniques to enable simultaneous translation. For example, (Ma et al., 2019) implements a wait-k model and utilizes meta-learning to address data scarcity, while (Zhang et al., 2022b) employs a wait-info model that incorporates information entropy from both the original text and the translation into the model. Additionally, (Liu et al., 2020) utilizes a unidirectional encoder with monotonic cross-attention to constrain dependence on future context.

In addition, some research has focused on detecting stable hypotheses. For instance, (Liu et al., 2020) proposed the Hold-n strategy, 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. Additionally, like the LA-n strategy, (Nguyen et al., 2021) developed the SP-n strategy, which identifies the longest common prefix among all items in the beam of a chunk. Our work directly addresses this issue.

3 Methods

Figure 1 illustrates our framework.

3.1 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 four different decoding strategies: "ctc_greedy_search", "ctc_beam_search", "attention_decoding", and "attention_rescoring". The CTC decoder, with argmax decoding, guarantees that the tokens decoded in previous chunks are unaltered, leading to a smooth streaming experience. The attention decoder generates output token by token and also has the ability to re-score CTC generated texts using prefix beam search in the event of multiple candidate proposals.

After building on our findings from last year, we have discovered that U2 offers stability and robustness in predicting audio without real utterances. This improvement is due to the model's training strategy, specifically the use of dynamic chunk training. In our current work, we have further improved the performance of the model by breaking the chunk-based attention approach and employing the "attention_rescoring" decoding strategy.

3.2 MT

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

Multilingual Translation (Johnson et al., 2017) has proposed a simple solution for translating multiple languages using a single neural machine translation model with no need to alter the model architecture. The proposed technique involves inserting an artificial token at the start of the input sentence to specify the target language. Furthermore, all languages use the same vocabulary, eliminating the need to add additional parameters. In this study, En-De/ZH/JA data was combined and jointly trained, demonstrating that a multilingual model can significantly enhance translation performance.

Data diversification Data diversification (Nguyen et al., 2020) is an effective strategy to improve the performance of NMT. This technique involves utilizing predictions from multiple forward and backward models and then combining the results with raw data to train the final NMT model. Unlike other methods such as knowledge distillation and dual learning, data diversification does not require additional monolingual data and can be used with any type of NMT model. Additionally, this strategy is more efficient and exhibits a strong correlation with model integration.

Forward translation Forward translation (Wu et al., 2019) refers to using monolingual data in the source language to generate synthetic data through beam search decoding. This synthetic data is then added to the training data in order to increase its size. While forward translation alone may not yield optimal results, when combined with a back translation strategy, it can enhance performance more effectively than back translation alone. In this work, we use only the forward model to create synthetic data and add the data to the original parallel corpora.

Domain Fine-tuning Previous studies have shown that fine-tuning a model with in-domain data can significantly enhance its performance. We hypothesize that there are domain-like distinctions between ASR-generated results and actual text. To further improve the performance, we use the generation from a well-trained ASR model to replace source-side text in the training corpus data. This fine-tuning approach enables us to achieve further improvements in the MT model.

3.3 Onlinization

Incremental Decoding Translation tasks may require reordering or additional information that is not apparent until the end of the source utterance, depending on the language pair. In offline settings, processing the entire utterance at once produces the highest-quality results. However, this approach also leads to significant latency in online mode. One possible solution to reduce latency is to divide the source utterance into smaller parts and translate each one separately.

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 predictions 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.

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. Their methods, such as LA-n (Liu et al., 2020) and SP-n (Nguyen et al., 2021), are designed for use in end-to-end systems that search for a shared prefix among the hypotheses generated from different chunk inputs. In contrast, our approach operates within a cascaded system that processes the same chunk input.

We can denote the MT and ASR generating functions as G and F respectively. Let $F_{i,n}^{C}$ 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:

Model	Language Pair	Lantency	BLEU	AL	AP	DAL
IWSLT22 Best System	EN-DE	Low	26.82	0.96	0.77	2.07
		Medium	31.47	1.93	0.86	2.96
		High	32.87	3.66	0.96	4.45
Our System	EN-DE	-	**33.54**	**1.88**	0.83	2.84
IWSLT22 Best System	EN-JA	Low	16.92	2.46	0.9	3.22
		Medium	16.94	3.77	0.97	4.29
		High	16.91	4.13	0.98	4.53
Our System	EN-JA	-	**17.89**	**1.98**	0.83	2.89
IWSLT22 Best System	EN-ZH	Low	25.87	1.99	0.87	3.35
		Medium	26.21	2.97	0.94	4.16
		High	26.46	3.97	0.98	4.62
Our System	EN-ZH	-	**27.23**	**1.98**	0.83	2.89

Table 1: Final systems results

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

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

4 Experiments Setup

4.1 ASR

Model 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 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., 2022a).

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.

Training 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.

4.2 MT

Model 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, $pre\_ln = True$.

Dataset 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.

Training 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 $\beta1$ = 0.9 and $\beta2$ = 0.98. During the inference phase, a beam size of 8 is used. The length penalties are set to 1.0.

5 Results

From Table 1, we can see that the our systems work well on various language pairs. And our systems even beat the best IWSLT 22 systems under higher latency.

Language Pair	Model	BLEU
En-DE	Offline	35.23
-	Simul	33.54
En-JA	Offline	19.45
-	Simul	17.89
En-ZH	Offline	27.93
-	Simul	27.23

Table 2: Comparison to offline system

Previous research has shown that the quality of simultaneous translation can now match or even surpass that of offline systems. However, in our current study, we first established a new baseline for the offline system. Furthermore, we found that there is still a difference of 1-2 BLEU between simultaneous translation and offline translation, see Table 2.

5.1 Ablation Study on different ASR decoding strategies

Language Pair	Decoding strategies	BLEU
En-DE	ctc_beam_search	32.88
En-JA	ctc_beam_search	16.56
En-ZH	ctc_beam_search	26.47
En-DE	attention_rescoring	33.54
En-JA	attention_rescoring	17.89
En-ZH	attention_rescoring	27.23

Table 3: Ablation Study on different ASR decoding strategies

The decoding strategy of "attention_rescoring" involves using a decoder to re-rank the results based on the decoding output of "ctc_beam_search". As a result, "attention_rescoring" can obtain better ASR results. Table 3 demonstrates that a better ASR decoding strategy can lead to overall better quality results for the system.

5.2 Ablation Study on MT training strategies

Training strategies	BLEU
Baseline	33.54
- Domain Fine-tuning	27.87
- Forward Translation	25.49
- Multiligual Translation	23.76

Table 4: Ablation Study on MT training strategies for EN-DE direction

In the field of machine translation, Domain Fine-tuning, Forward Translation, and Multiligual Translation are frequently employed methods to enhance translation quality. It is evident from Table 4 that these training strategies can effectively improve the overall quality of the system.

6 Conclusion

In this paper, we report on our work in the IWSLT 2023 simultaneous speech-to-text translation evaluation. We propose an onlinization strategy that can be applied to cascaded systems and demonstrate its effectiveness in three language directions. Our approach is simple and efficient, with ASR and MT modules that can be optimized independently. Our cascade simultaneous system achieves results that are comparable to offline systems. In the future, we plan to further explore the direction of 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, Yinglu Li, Minghan Wang, Xiaosong Qiao, Yuxia Wang, Hengchao Shang, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022. The hw-tsc's speech to speech translation system for IWSLT 2022 evaluation. 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 293–297. 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.

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.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 3025–3036. Association for Computational Linguistics.

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.

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.

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.

Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022a. The hw-tsc's simultaneous speech translation system for IWSLT 2022 evaluation. 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 247–254. Association for Computational Linguistics.

Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022b. The hw-tsc's offline speech translation system for IWSLT 2022 evaluation. 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 239–246. Association for Computational Linguistics.

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. 2022a. 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.

Shaolei Zhang, Shoutao Guo, and Yang Feng. 2022b. Wait-info policy: Balancing source and target at information level for simultaneous machine translation. In Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 2249–2263. Association for Computational Linguistics.