SimulSeamless: FBK at IWSLT 2024 Simultaneous Speech Translation

Sara Papi and **Marco Gaido** and **Matteo Negri** and **Luisa Bentivogli**
Fondazione Bruno Kessler, Italy
{mgaido,spapi,negri,bentivo}@fbk.eu

Abstract

This paper describes the FBK's participation in the Simultaneous Translation Evaluation Campaign at IWSLT 2024. For this year's submission in the speech-to-text translation (ST) sub-track, we propose **SimulSeamless**, which is realized by combining AlignAtt and SeamlessM4T in its medium configuration. The SeamlessM4T model is used "off-the-shelf" and its simultaneous inference is enabled through the adoption of AlignAtt, a SimulST policy based on cross-attention that can be applied without any retraining or adaptation of the underlying model for the simultaneous task. We participated in all the Shared Task languages (English→ {German, Japanese, Chinese}, and Czech→English), achieving acceptable or even better results compared to last year's submissions. SimulSeamless, covering more than 143 source languages and 200 target languages, is released at https://github.com/hlt-mt/FBK-fairseq/.

1 Introduction

Simultaneous speech-to-text translation (SimulST) is the task in which a model has to provide a textual translation into the target language while continuously receiving an incremental speech input in the source language.

SimulST poses additional difficulties to standard offline ST, as it has to find the optimal balance between translation quality and output latency, which is the time delay between an utterance being spoken and the corresponding translation being emitted. This balance – often referred to as "quality-latency tradeoff" – depends on the application scenario (Fantinuoli and Prandi, 2021), which can span many domains such as online meetings, lectures, conference talks, and live shows.

Due to the growing interest in SimulST technologies, this task has been included in the IWSLT Evaluation Campaigns[1] since 2020. The increasing interest has led to numerous direct and cascade models participating in the challenge every year (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023), all vying for the title of the best approach to realize a SimulST system from scratch. More recently, the practice of using models without ad-hoc training for the simultaneous scenario has become widespread (Polák et al., 2022; Gaido et al., 2022; Papi et al., 2023a; Polák et al., 2023; Yan et al., 2023; Huang et al., 2023), demonstrating that competitive or even superior results can be achieved compared to systems specifically tailored for SimulST (Papi et al., 2022a). Among the strategies used to repurpose standard (offline) ST models for SimulST (Liu et al., 2020; Papi et al., 2022a, 2023c), AlignAtt (Papi et al., 2023b) emerged as the best one, achieving new state-of-the-art results. AlignAtt exploits speech-translations alignments based on cross-attention scores to guide the simultaneous inference, overcoming the limitations of the previous approach relying on attention (Papi et al., 2023c).

Alongside the increased interest in the SimulST task, especially during the last year, we have witnessed an explosion in the use of large models (Latif et al., 2023), including speech foundation models (Radford et al., 2023; Pratap et al., 2023; Barrault et al., 2023a; Zhang et al., 2023). These models are now commonly used alone or in combination with large language models (Gaido et al., 2024) for generic ST tasks. Among these, SeamlessM4T (Barrault et al., 2023a) has emerged as one of the most promising multimodal and multilingual models, covering more than 143 source languages and 200 target languages.

For this year's submission to the IWSLT Evaluation Campaign on Simultaneous Translation, we, therefore, propose to combine the best of both

[1] https://iwslt.org/

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 120–127
August 15-16, 2024

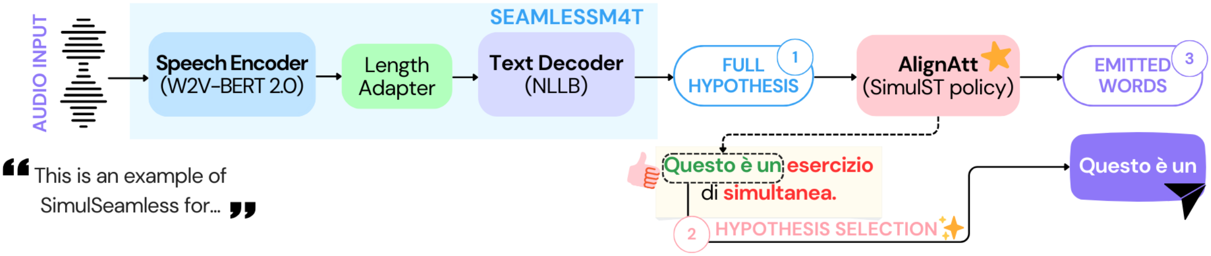


Figure 1: Representation of the SeamlessM4T model combined with AlignAtt: SimulSeamless.

worlds to obtain a multilingual model without any training or adaptation for the SimulST task. This results in **SimulSeamless**, consisting of the SeamlessM4T model used "off-the-shelf" repurposed for simultaneous inference using AlignAtt.

From empirical results on the task, we show that SimulSeamless can achieve acceptable or even better results compared to last year's participants, despite not being retrained or fine-tuned either for the simultaneous task or on paired data in the evaluated languages. Moreover, SimulSeamless is a generic multilingual model that can be used for any allowed translation direction supported by the underlying SeamlessM4T model, covering more than 143 source languages and 200 target languages. The code is released under the Apache 2.0 Licence at `https://github.com/hlt-mt/FBK-fairseq/blob/master/fbk_works/SIMULSEAMLESS.md`.

2 SimulSeamless

Similarly to previous years (Gaido et al., 2022; Papi et al., 2023a), we participated in the Simultaneous Translation evaluation campaign, focusing on the speech-to-text translation sub-track. For this year's submission, we opted for the use of the new SeamlessM4T model, which is allowed for the task,[2] as the underlying model of the SimulST policy AlignAtt. This policy can be applied to any standard (i.e., offline-trained) model without the need for retraining or adaptation.

In the following, both these elements and their combination are explained in detail.

SeamlessM4T. SeamlessM4T (Barrault et al., 2023a) (or Massively Multilingual & Multimodal Machine Translation) is a family of models based on pre-trained models including W2V BERT 2.0, and NLLB (Costa-jussà et al., 2022), whose encoder and decoder respectively are used for the speech-to-text modality. W2V-BERT is a Conformer-based model (Gulati et al., 2020) composed of 24 layers, with a total of ∼600M parameters, and trained on 1 million hours of open speech audio data to learn self-supervised speech representations. It processes the audio features obtained by applying 80-dimensional Mel filterbanks to the audio waveform. The W2V-BERT encoder is followed by a Length Adapter based on a modified version of the M-adaptor (Zhao et al., 2022), which is a Transformer-based model (Vaswani et al., 2017) that is in charge of compressing the speech representation (by a factor of 8) through attention pooling. The compressed input representations are then fed to the NLLB decoder, in its 1.3B parameters configuration, to produce the translations. The final model was obtained after training on both manual and automatically aligned speech translation data with a total of 406,000 hours.

AlignAtt. AlignAtt (Papi et al., 2023b) is a SimulST policy that relies on cross-attention to make decisions about whether to emit translated words or wait for additional information in the simultaneous scenario. At each time step, the cross-attention scores are exploited to obtain audio-translation alignments by uniquely assigning the predicted words to the audio frames (encoder states) having the maximum attention score. Then, it is checked, for each word, if it has been aligned with one of the last f frames, which is the parameter handling the latency of the model. If this is true, the emission is stopped, otherwise, the next word is evaluated. The idea behind AlignAtt is that, if a word is aligned with one of the last received audio frames, the encoded information could be unstable and/or not sufficient to reliably predict that word. Conversely, if a word mostly attends to a more stable and earliest-received encoded information, it can be safely predicted. With this formulation, AlignAtt simplifies the previous EDAtt policy (Papi et al., 2023c) by eliminating the dependency on additional hyper-parameters while achieving com-

[2] `https://iwslt.org/2024/simultaneous`

petitive or even better results.

SeamlessM4T + AlignAtt = SimulSeamless. Since AlignAtt is applicable to any standard ST models without the need for re-training or adaptation, we chose to apply it directly to the SeamlessM4T model in its medium configuration, realizing **SimulSeamless**. This solution is completely different from SeamlessStreaming (Barrault et al., 2023b), which is obtained through an expensive ad-hoc finetuning of the Seamless model for the simultaneous task based on EMMA – efficient monotonic multi-head attention (Ma et al., 2023). Since SeamlessM4T already covers all the languages evaluated in the Simultaneous track, the model is used completely "off-the-shelf". The SimulSeamless model is shown in Figure 1.

3 Experimental Settings

We used the available checkpoint of the SeamlessM4T model provided on HuggingFace in its "medium" configuration,[3] with a total of 1.2B parameters.

The results are reported on the benchmarks used for the submission, which is MuST-C (Cattoni et al., 2021) v2.0 tst-COMMON for en-{de, ja, zh}, and the dev set provided for the task for cs-en. The scores are computed using the SimulEval toolkit (Ma et al., 2020).[4] Translation quality is evaluated using BLEU score with sacreBLEU (Post, 2018)[5]. Latency is reported using Average Lagging (AL) (Ma et al., 2019) since it is the metric used for the final scoring. Length Adaptive Average Lagging (LAAL) (Papi et al., 2022b) and Average Token Delay (ATD) (Kano et al., 2022) are also evaluated and included in the final results since they are official metrics reported for the task.[6] Both latency and BLEU scores are computed at the character level for Chinese and Japanese while the standard `13a` tokenizer is used for sacreBLEU, and word-level latency is computed for the other languages. Additionally, computationally aware metrics are presented to account for the real elapsed time, which also considers the computational cost of running the underlying model. The inference was run using a single GPU NVIDIA V100 with 16GB of RAM.

[3]https://huggingface.co/facebook/seamless-m4t-medium

[4]We used the `f1f5b9a` commit that is the last version with the remove evaluation working, which is needed to run SimulEval using Docker containers.

[5]Version 2.4.0.

[6]https://iwslt.org/2024/simultaneous

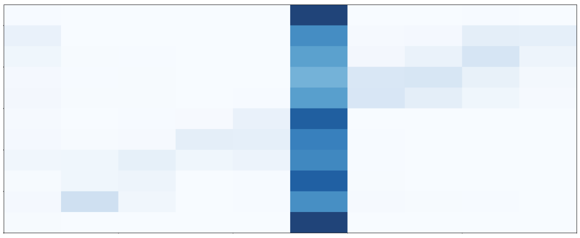
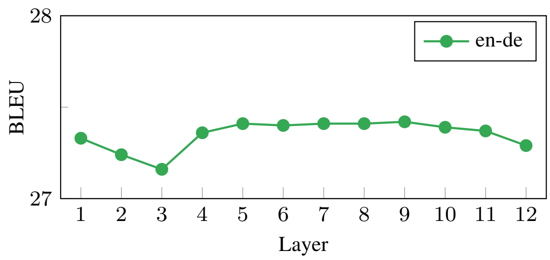
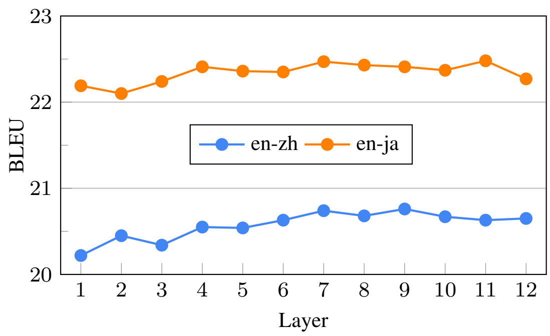


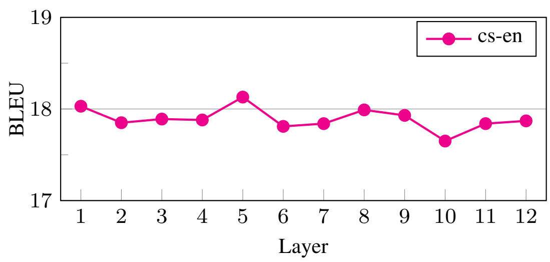
Figure 2: Example of skewed cross-attention scores representation towards some frames.



(a) English to German



(b) English to Chinese and Japanese



(c) Czech to English

Figure 3: Translation quality (BLEU↑) scores of SimulSeamless on MuST-C v2.0 tst-COMMON for English (en) to German (de), Japanese (ja), and Chinese (zh), and on the IWSLT 2024 dev set for Czech (cs) to English by varying the decoder layer from which cross-attention scores are extracted from.

For the AlignAtt policy, we set the size of the speech chunk processed by the model at each time step to 1s for English to German and Czech to English, 800ms for English to Chinese, and 400ms for English to Japanese. To achieve latency close to an AL of 2s required for the submission, we set

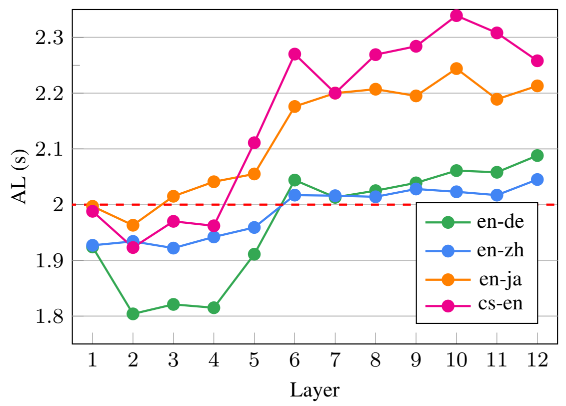


Figure 4: Latency (AL↓) scores of SimulSeamless on MuST-C v2.0 tst-COMMON for English (en) to German (de), Japanese (ja), and Chinese (zh), and on the IWSLT 2024 dev set for Czech (cs) to English by varying the decoder layer from which cross-attention scores are extracted from.

the hyper-parameter handling the latency f to 1 for en-ja and en-zh, 6 for en-de, and 9 for cs-en. The cross-attention scores are normalized frame-wise before applying AlignAtt to avoid the cross-attention weights being skewed to some frame representation, as shown in Figure 2.

4 Results

4.1 Submission Selection

For selecting the best setting, we analyzed the performance by varying the layer from which cross-attention scores are extracted since simply averaging them across layers led to worse results, as also already found in (Papi et al., 2023c). The layer-wise quality results are shown in Figure 3 while layer-wise latency results close to AL=2s are shown in Figure 4.

It can be seen from the Layer-AL(s) curves (Figure 4) that Layer 5 represents a threshold layer starting from which the latency increases significantly without, however, similar significant quality improvements in terms of BLEU (Figure 3). The only acceptable layers to achieve an AL≤2s for en-ja are layers 1 and 2 while this set is extended to layer 4 for cs-en, and up to layer 5 for en-de and en-zh. Among the two admissible layers for en-ja, we chose for the final submission the one maximizing the quality, which is Layer 1. For en-zh, we followed a similar approach by choosing Layer 4, which achieves the highest BLEU score with an admissible latency. The choice of Layer 4 is also maintained for en-de and cs-en since we found that is the layer achieving the best quality-latency tradeoff between BLEU and AL.

4.2 Comparison with Last Year's Participants

In Table 1, we report the scores for the final submission for each language pair, including LAAL and ATD latency metrics and their corresponding computationally aware scores. SimulSeamless is compared with all the participants of last year: CMU (Yan et al., 2023), CUNI-KIT (Polák et al., 2023), FBK (Papi et al., 2023a), HW-TSC (Guo et al., 2023), NAIST (Fukuda et al., 2023), and XIAOMI (Huang et al., 2023). Comparisons are not reported for cs-en since it is a new language direction for the task.

First, it can be noticed that SimulSeamless achieves the best translation quality and, in general, the best quality-latency trade-off for en-ja. Conversely, it struggles to achieve very competitive results in en-de and, especially, in en-zh. However, it is important to notice that SimulSeamless is the only model that has not been fine-tuned on the IWSLT-allowed data for the task, which include the MuST-C v2.0 training set. Therefore, it is a more generic and multilingual system covering more than 143 source languages and 200 target languages.[7]

Furthermore, an overlap has been identified between the MuST-C tst-COMMON and the ST-TED dataset (Zhang and Ao, 2022), which was allowed for last year's task. Some participants, unaware of this issue, employed the ST-TED dataset (e.g., CUNI-KIT and XIAOMI). Therefore, the results achieved by last year's submissions on the MuST-C tst-COMMON may not be entirely reliable. In addition, it has been recently found another possible overlap with TED2020, which may invalidate other scores.[8]

In conclusion, SimulSeamless allows for acceptable or even better results compared to last year's participants in the SimulST Evaluation Campaign while being generic and potentially applicable to all translation directions supported by the underlying SeamlessM4T model without any retraining or adaptation.

[7]We are not able to exclude that MuST-C has been used for training the "off-the-shelf" SeamlessM4T but no ad-hoc fine-tuning on the data and/or language pairs has been performed for our participation.

[8]Unaware of this overlap, participations from CMU and HW-TSC used this dataset.

Lang. Pair	Model	BLEU ↑	AL ↓	LAAL ↓	ATD ↓
en-de	CMU†	30.4	1.92	1.99	-
	CUNI-KIT†	31.4	1.955 (3.072)	-	-
	FBK†	30.70	1.888 (2.939)	2.069 (3.052)	1.797 (2.364)
	HW-TSC‡	33.54	1.88	-	-
	NAIST	29.98	1.964	2.173	1.894
	SimulSeamless†	27.37	1.815 (3.012)	1.993 (3.137)	1.778 (2.353)
en-ja	NAIST	15.32	1.974	2.291	0.548
	CUNI-KIT†	15.3	1.982 (3.489)	-	-
	HW-TSC‡	17.89	1.98	-	-
	SimulSeamless†	22.19	1.997 (4.018)	2.137 (4.272)	0.580 (2.728)
en-zh	NAIST	22.11	1.471	1.907	0.668
	CUNI-KIT†	26.6	1.987 (3.508)	-	-
	HW-TSC‡	27.23	1.98	-	-
	XIAOMI†	26.59	1.966	-	-
	SimulSeamless†	20.56	1.942 (3.388)	2.080 (3.465)	0.765 (1.933)
cs-en	**SimulSeamless**†	18.03	1.988 (3.755)	2368 (3.999)	2.778 (3.399)

Table 1: Results on the MuST-C v2.0 tst-COMMON (for en-{de, ja, zh}) and IWSLT 2024 dev (for cs-en) considering BLEU and all the latency metrics (in seconds) reported for the task. Results in brackets are computationally aware but computed with different environments between systems. † indicates systems trained offline and tested in simultaneous. ‡ indicates cascade systems.

5 Conclusions

We introduced FBK's system designed for participation in the IWSLT 2024 Evaluation Campaigns in Simultaneous Translation and, specifically, the speech-to-text sub-track (SimulST). Our submission is characterized by the "off-the-self" use of the SeamlessM4T model for direct speech translation, repurposed for the simultaneous scenario by means of AlignAtt. AlignAtt is a SimulST policy that leverages cross-attention scores to guide simultaneous inference without any further modification or adaptation of the underlying model. The combination of SeamlessM4T and AlignAtt results in SimulSeamless, which supports all translation pairs of the Evaluation Campaign (English to German, Japanese, and Chinese, and Czech to English). SimulSeamless, to be released upon paper acceptance, achieves acceptable or even superior results compared to last year's participants. Moreover, it can be used for any language pairs enabled by the underlying SeamlessM4T model, potentially covering more than 143 source languages and 200 target languages.

Acknowledgments

The work presented in this paper is funded by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People), and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, 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, Xutai 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 Kr. Ojha, 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)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online).

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online).

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023a. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023b. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics.

Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2023. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 330–340, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022. Efficient yet competitive speech translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? *arXiv preprint arXiv:2402.12025*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

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.

Wuwei Huang, Mengge Liu, Xiang Li, Yanzhi Tian, Fengyu Yang, Wen Zhang, Jian Luan, Bin Wang, Yuhang Guo, and Jinsong Su. 2023. The xiaomi AI lab's speech translation systems for IWSLT 2023 offline task, simultaneous task and speech-to-speech task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 411–419, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2022. Average token delay: A latency metric for simultaneous translation. *arXiv preprint arXiv:2211.13173*.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of Large Audio Models: A Survey and Outlook. *arXiv preprint arXiv:2308.12792*.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Proc. Interspeech 2020*, pages 3620–3624.

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 Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online.

Xutai Ma, Anna Sun, Siqi Ouyang, Hirofumi Inaguma, and Paden Tomasello. 2023. Efficient monotonic multihead attention. *arXiv preprint arXiv:2312.04515*.

Sara Papi, Marco Gaido, and Matteo Negri. 2023a. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022a. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022b. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online.

Sara Papi, Matteo Negri, and Marco Turchi. 2023b. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Proc. of Interspeech 2023*, Dublin, Ireland.

Sara Papi, Matteo Negri, and Marco Turchi. 2023c. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics.

Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. 2023. Towards efficient simultaneous speech translation: CUNI-KIT system for simultaneous track at IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 389–396, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej 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 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling Speech Technology to 1,000+ Languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Brian Yan, Jiatong Shi, Soumi Maiti, William Chen, Xinjian Li, Yifan Peng, Siddhant Arora, and Shinji Watanabe. 2023. CMU's IWSLT 2023 simultaneous speech translation system. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 235–240, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Ziqiang Zhang and Junyi Ao. 2022. The YiTrans speech translation system for IWSLT 2022 offline shared task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 158–168, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jinming Zhao, Hao Yang, Gholamreza Haffari, and Ehsan Shareghi. 2022. M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation. In *Proc. Interspeech 2022*, pages 111–115.