

# MLLP-VRAIN UPV system for the IWSLT 2025 Simultaneous Speech Translation Translation task

Jorge Iranzo-Sánchez\*<sup>◇</sup> and Javier Iranzo-Sánchez<sup>◇</sup>  
and Adrià Giménez<sup>†</sup> and Jorge Civera<sup>◇</sup> and Alfons Juan<sup>◇</sup>

<sup>◇</sup>Machine Learning and Language Processing, VRAIN, Universitat Politècnica de València

<sup>†</sup>Departament d'Informàtica, Escola Tècnica Superior d'Enginyeria, Universitat de València  
{jorirsan, jairsan, jorcisai, ajuanci}@upv.es, adria.gimenez@uv.es

## Abstract

This work describes the participation of the MLLP-VRAIN research group in the shared task of the IWSLT 2025 Simultaneous Speech Translation track. Our submission addresses the unique challenges of real-time translation of long-form speech by developing a modular cascade system that adapts strong pre-trained models to streaming scenarios. We combine Whisper Large-V3-Turbo for ASR with the multilingual NLLB-3.3B model for MT, implementing lightweight adaptation techniques rather than training new end-to-end models from scratch. Our approach employs document-level adaptation with prefix training to enhance the MT model's ability to handle incomplete inputs, while incorporating adaptive emission policies including a wait- $k$  strategy and RALCP for managing the translation stream. Specialized buffer management techniques and segmentation strategies ensure coherent translations across long audio sequences. Experimental results on the ACL60/60 dataset demonstrate that our system achieves a favorable balance between translation quality and latency, with a BLEU score of 31.96 and non-computational-aware StreamLAAL latency of 2.94 seconds. Our final model achieves a preliminary score on the official test set (IWSLT25Instruct) of 29.8 BLEU. Our work demonstrates that carefully adapted pre-trained components can create effective simultaneous translation systems for long-form content without requiring extensive in-domain parallel data or specialized end-to-end training.

## 1 Introduction

In this paper we describe the participation of the MLLP-VRAIN research group in the shared tasks of the 22th International Conference on Spoken Language Translation (IWSLT) (Abdulmumin et al., 2025). We participated on the Simultaneous Speech Translation (SimulST) task in the English to German direction. Compared to other years,

two aspects were changed in the shared task which guided the construction of our system: The usage of pretrained open weight models and the evaluation of long-form audio. Our participation this year was an attempt of creating a production ready model based on the minimal adaptation of offline ASR and MT (Papi et al., 2022). Recent years have seen a rise in the usage of end-to-end approaches<sup>1</sup> which, in theory, can offer a better integration and can avoid error compounding between ASR and MT components. However, they typically require large quantities of parallel speech-to-text translation data, which is often scarce and costly to obtain. Cascade systems, on the other hand, while they are more data-efficient due to the abundance of separate ASR and MT training resources, may suffer from error propagation and lack of joint optimization. Nevertheless, recent shared tasks and evaluation campaigns continue to show that cascade systems generally achieve superior performance over current end-to-end alternatives (Salesky et al., 2023b, 2024). As such, we model our system based on the cascaded approach, in which we take a special keen interest due to its inherent modularity and easier reuse of strong pre-trained components. Figure 1 shows the overall architecture of our system.

## 2 System Architecture

### 2.1 ASR system

For the choice of the ASR components, we select our model based on the results of public ASR systems on common benchmarks available at the Hugging Face Open ASR Leaderboard (Srivastav et al., 2023). After some initial tests and taking into account our computing limitations, we selected Whisper-Large-V3-Turbo (Radford et al., 2023)<sup>2</sup>

<sup>1</sup>As defined by IWSLT <https://iwslt.org/2025/offline#evaluation-conditions>.

<sup>2</sup><https://huggingface.co/openai/whisper-large-v3-turbo>

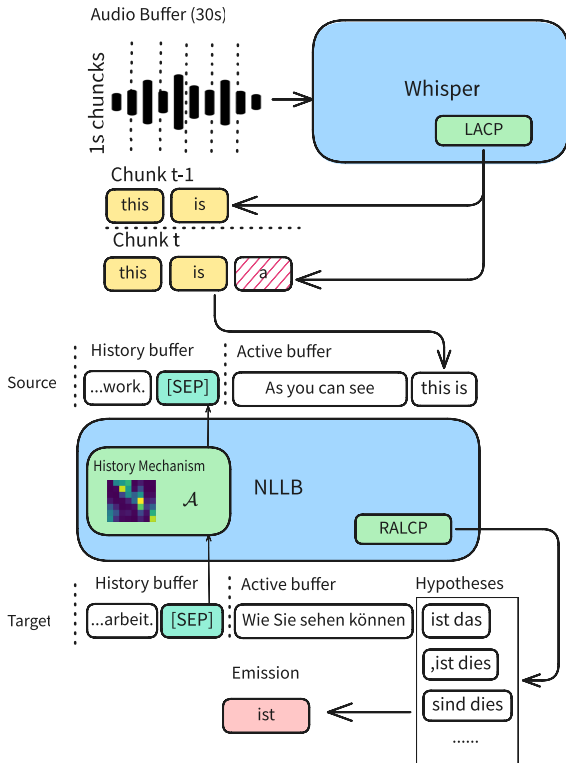


Figure 1: System diagram of our cascaded system for the SimulST track.

as our final ASR model. We use the model as it is, and we do not make usage of any of the English audio data provided by the organizers for finetuning of the model. There are two reasons for this decision. First, when looking at the provided datasets, we conjecture that the Whisper model has probably have already seen this data; and second, that we fear that we may end up lowering the performance of the system as there is a domain mismatch between the provided datasets and evaluation set, with the latter being scientific talks of the ACL.

The adaptation for streaming is done in a similar way to the one described in Macháček et al. (2023), where a Longest Common Prefix policy along some heuristics are combined for the usage in a streaming fashion. Inference is done via the Faster-Whisper library<sup>3</sup>. We select a maximum audio buffer of 30 seconds and a minimum chunk size of one second and deactivate the usage of VAD filtering. The audio buffer is cleaned when a end of sentence is detected by external sentence splitter or the 30 seconds window is full. The audio stream is then updated accordingly to the timestamps obtained by the typical DTW procedure used in Whisper.

During development we detected that this base

<sup>3</sup><https://github.com/SYSTRAN/faster-whisper>

system had sometimes very unreliable behaviour due to latency spikes derived from indecisions of the LCP policy on the punctuation, casing and styling of some words. To alleviate this, we relaxed the LCP policy so that it is done on lower cased input, with no punctuation signs, and with the additional constraint of a threshold by Levenshtein Distance to consider if a word in the prefix is sufficiently different. We select a Levenshtein Distance threshold of two so that if the distance of the words to be checked is less or equals to it, then they are to be considered the same by the policy.

The ASR system obtains a WER of 8.54% on the ACL60/60 (Salesky et al., 2023a) development set.

## 2.2 MT system

During recent years a series of works have appeared exploring realistic scenarios for SimulST where models are evaluated on long-form speech scenarios. (Schneider and Waibel, 2020; Sen et al., 2022; Papi et al., 2024; Polák and Bojar, 2023; Iranzo-Sánchez et al., 2024; Ouyang et al., 2025). Taking this into consideration, for our MT model, we adapt an offline MT model to streaming through a lightweight procedure inspired to that of Iranzo-Sánchez et al. (2024), which presents an easy to adapt pipeline for the creation of an MT component in SimulST system which we further adapt and simplify it to our given conditions. Overall, through training with prefix-training (Niehues et al., 2018) and document level metadata, the model is able to learn to work with an incomplete stream and control a history stream (Iranzo-Sánchez et al., 2022) by emitting a sentinel token [SEP] which then serves to call a lightweight log-linear model to take into account which part of the stream has already been translated. When a certain threshold of size in the buffer reached, the oldest pairs of identified segments can easily be discarded without any fear of possible mismatches in the stream.

**Model:** Instead of training from scratch, we make use of the multilingual system NLLB-3.3B (Team, 2024). We also explored the usage of MADLAD-400 (Kudugunta et al., 2023), but discarded it after founding BLEU scores of over 95 points when evaluation of the system in an offline scenario for the ACL60/60 dataset, which indicated a possible data contamination of the test set. We also tested some LLMs, but found early results too unsatisfactory for our computing budget.

**History Mechanism:** Inspired by the usage of attention maps of Papi et al. (2023), for the log-feature model, instead of the described reverse translation model of the original paper, we found that a single feature which looks at the most attended position of the previous token before the sentinel [SEP] token to be more simpler and effective to determine the segmentation position of the source stream. To be more precise, the position of the last source word  $\hat{a}$  to be moved to the streaming history buffer for the current active source and target chunks  $x$  and  $\hat{y}$  is

$$\hat{a} = \underset{i}{\operatorname{argmax}} \mathcal{A}(x_i, y_{[SEP]-1}) \quad (1)$$

with  $\mathcal{A}$  being the attention score function and  $i$  indicates the position of source word  $x$  in the active chunk.

**Data and Training:** For the document level data, we take the available News Comentary<sup>4</sup> and Europarl<sup>5</sup> datasets with their reconstructed document level information extracted from Paradocs (Wicks et al., 2024) for a total of 36225 documents. While not in the target domain of ACL talks, we hope that our adaptation is able to move the model from offline inference domain to simultaneous with the usage of history context. Additionally, we want to adapt the multilingual model so that it can better focus on the English to German direction. Prefix augmented data with document level information is created as in the original work, but is dynamically generated at training time. For each document, which represents a data sample, we select a sentence and randomly prepend from 1 up to 10 of the previous phrases of the document with the corresponding sentinel token [SEP]. Then, for the corresponding phrase we create the prefix by taking into account the ratio of the length of the active source and target phrase. During training, the usage of prefix training is triggered at a rate of 50%. As for the training procedure, we make use DoRA (Mao et al., 2024) to fine-tune our MT model. Training hyperparameters are shown in Appendix A.

**Policy:** For our policy in the MT component, we make use of RALCP (Wang et al., 2024) using the hypotheses of the system beam search in combination with a *wait-k* policy (Ma et al., 2019), the latter

<sup>4</sup><https://data.statmt.org/news-commentary/v18/training/>

<sup>5</sup><https://www.statmt.org/europarl/v10/training/europarl-v10.de-en.tsv.gz>

being active only during the beginning of a new phrase (that is after emitting a [SEP] token) to prevent system hallucination. We also clean “invalid” empty beam hypotheses, which were very frequent on the baseline model, adjusting the RALCP  $\lambda$  accordingly so that the ratio of hypotheses and  $\lambda$  stays roughly the same as before filtering.

The final peak memory usage during inference of both ASR and MT models combined is of 20GB of VRAM<sup>6</sup>, with values fluctuating between 1-3GiB depending on the current audio and translation history buffers. We leave further optimization for future work.

### 3 Evaluation

For our baselines, we searched for any other public simultaneous speech translation systems capable of long form translation. However, we only found the system described in Papi et al. (2024) available, and in this case, we observed that the system had strong hallucinations that ended up in an unrecoverable state for long enough audios. Due to this, we instead build a “naive” baseline for our system which make use of the plain translation system without any additional training. For controlling the history buffers, we simply remove from the source and target text history buffers after a maximum number of words is reached in each buffer. In practice, we surpassingly found that while the system will end up with slight mismatches between the source and target stream at the head position, systems are still able to work in a streaming scenario. We additionally add the offline inference of the model before and after fine-tuning and the best baseline from the IWSLT organizers that achieved a similar latency-quality tradeoff matching our own.<sup>7</sup> Best hyperparameters with the optimal quality/latency are shown in Appendix A.

Table 1 shows the results of the baseline and our adapted model, the stream LAAL, both computationally an computationally aware, BLEU (Papineni et al., 2002) as calculated with SacreBLEU (Post, 2018)<sup>8</sup> and the COMET-22 (Rei et al., 2022)<sup>9</sup>. We follow the recommendations of Zouhar

<sup>6</sup>On a Nvidia GTX 4090 with a Intel(R) Core(TM) i9-10920X CPU @3.50GHz

<sup>7</sup>Organizer baseline results extracted from [https://github.com/pe-trik/iwslt25-baselines/tree/master/experiments/acl6060\\_dev/de/cascade](https://github.com/pe-trik/iwslt25-baselines/tree/master/experiments/acl6060_dev/de/cascade)

<sup>8</sup>Python3.12.9|BLEU|nrefs:1|case:mixed|eff:n|tok:13|smooth:expl|version:2.5.1

<sup>9</sup>Python3.12.9|Comet2.2.6|fp16|Unbabel/wmt22-comet-dalr1

	StreamLAAL			
	BLEU	COMET	NCA	CA
Offline	43.12	0.833	—	—
Offline <sup>(A)</sup>	41.48	0.836	—	—
Baseline <sup>iwslt</sup>	25.47	—	3.67	—
Baseline <sup>upv</sup>	26.10	0.642	3.61	4.35
Adapted	31.96	0.732	2.94	4.20

Table 1: Quality (BLEU, COMET $\uparrow$ ) and non-computational and computational aware (NCA/CA) latency (StreamLAAL (secs) $\downarrow$ ) results on the ACL60/60 development set. Offline models take the golden reference source text. The Offline<sup>(A)</sup> model refers to the results of our adapted model when doing offline inference. Offline models results are obtained given the golden source reference transcription.

et al. (2024) and set the COMET score to 0 for samples where the target translation is empty after re-segmentation with mWERSegmenter (Matusov et al., 2005).

First of all, impact of the translation quality degradation in the offline mode seems to minimal, with deltas of 1.64 BLEU and 0.03 of COMET respectively. As for our baseline and adapted models, we can see that there is a significant quality degradation. We can attribute this to multiple factors, but we would like to highlight two aspects. First, is that we are comparing ourselves to offline systems that take the golden reference transcription, and thus, transcriptions errors from the ASR and possible resegmentation errors introduced by StreamLAAL are not taken into account in the evaluation of our offline baseline. The second factor which may explain this gap is the mere nature of the ACL60/60 dataset. We would like to highlight this one in particular since the translations were originally based on post-edits from offline translation models and thus more suited for the evaluation of offline speech translation compared to that of SimulST. Thus, we think that the resulting translations from our SimulST system, which should be more monotonous in nature compared to the offline baselines and ACL60/60 references, maybe be more penalized in a similar way to the observations of Doi et al. (2024).

In terms of baselines, we can see that our naive baseline slightly beats the organizers baseline in the selected quality-latency range. When comparing our baseline and adapted model, we can see a considerable increase in both quality and reduction

model	M	mdn	p90	p95	p99	max
NCA						
Baseline	3.61	2.96	6.57	9.14	13.51	16.59
Adapted	2.94	2.65	4.30	5.38	7.59	9.25
CA						
Baseline	4.35	3.62	7.65	10.51	14.58	18.52
Adapted	4.20	3.55	5.98	7.64	10.73	15.00

Table 2: StreamLAAL mean (M), median (mdn), percentiles 90%, 95% and 99%, and maximum value (in seconds) for the Baseline<sup>upv</sup> and Adapted system.

of latency. Our adapted model ends up scoring 31.96 BLEU and 0.732 COMET and StreamLAAL scores of 2.94 and 4.20 seconds depending on the computational awareness of the metric calculation. This places our model on the high latency regime as defined by the shared task description.

To better study the latency of our system, Table 2 shows the top percentiles of StreamLAAL as well as their medians and maximum recorded values. We can see how for all of these metrics our adapted system consistently beats our baseline and ensures a better performance on the worst case scenarios, with these delays being the more impactful for the end user of SimulST systems.

An observation that can be made in this table is that of the considerable increase of latency for the worst cases between the NCA and CA metrics. After investigating, we discovered the temperature fallback mechanism of Whisper to seem to cause this phenomena, resulting in some rare cases where latency spikes occur and can only be observed when taking computational costs into account. Despite this, we found that in practice the performance is really poor with this feature disabled. In general, we observed that the ASR emission policy highly influenced the system performance, with hyperparameter changes on the MT system having less of an overall impact.

Regarding the official test set (IWSLT25Instruct), preliminary results by the organizers indicate that our model achieved a final score of 29.8 BLEU.

## 4 Conclusions

In this paper we described our SimulST system for the IWSLT 2025 Simultaneous Speech Translation task. Preliminary results show that our cascaded



based system using Whisper and NLLB showed a good performance and achieved a good balance between translation quality and latency. We see how through adaptive policies and very computationally cheap adaptation a long-form speech SimulST system can be created from offline models. Future work could be expanded to see the robustness of this methodology, such as the usage of synthetic document level bitext data (Post and Junczys-Dowmunt, 2024) or speech data. Investigating more robust adaptive latency policies or techniques which better optimize ASR and MT components (Tran et al., 2022) while preserving the benefits of the cascaded approach could further greatly enhance the system performance. Also, a gap still exists compared to offline translation, which should be further explored in more detail.

Additionally, the usage of LLMs to serve as all in one transcriber, translator and re-scorer in a cascaded pipeline along their robustness and usage of long context shows promising results for their usage in long-form speech translation if computational costs can be taken into account.

## 5 Limitations

Due to time constraints, hyperparameters search for trade-offs between translation and latency of models was limited, as well as the tuning of the ASR system. In our participation, we restricted ourselves to the English to German direction, but we think that our approach could be generalized to the other language pairs in the competition. We hope to participate in future editions covering all language pairs available and expanding the breadth and scale of the studied models.

## Acknowledgments

The research leading to these results has received funding from EU4Health Programme 2021–2027 as part of Europe’s Beating Cancer Plan under Grant Agreements nos. 101056995 and 101129375; and from the Government of Spain’s grant PID2021-122443OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by “ERDF/EU”, and grant PDC2022-133049-I00 funded by MICIU/AEI/10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”. The authors gratefully acknowledge the financial support of Generalitat Valenciana under project IDIFEDER/2021/059.

## References

- Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połec, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maïke Züfle. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online). Association for Computational Linguistics. To appear.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. [8-bit optimizers via block-wise quantization](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Kosuke Doi, Yuka Ko, Mana Makinae, Katsuhito Sudoh, and Satoshi Nakamura. 2024. [Word order in English-Japanese simultaneous interpretation: Analyses and evaluation using chunk-wise monotonic translation](#). In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 254–264, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.
- Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2022. [From simultaneous to streaming machine translation by leveraging streaming history](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6972–6985, Dublin, Ireland. Association for Computational Linguistics.
- Javier Iranzo-Sánchez, Jorge Iranzo-Sánchez, Adrià Giménez, Jorge Civera, and Alfons Juan. 2024. [Segmentation-free streaming machine translation](#). *Transactions of the Association for Computational Linguistics*, 12:1104–1121.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- 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. Association for Computational Linguistics.
- Dominik Macháček, Raj Dabre, and Ondřej Bojar. 2023. [Turning whisper into real-time transcription system](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 17–24, Bali, Indonesia. Association for Computational Linguistics.
- Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. [Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11662–11675.
- Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. [Evaluating machine translation output with automatic sentence segmentation](#). In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.
- Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. [Low-latency neural speech translation](#). In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, pages 1293–1297.
- Siqi Ouyang, Xi Xu, and Lei Li. 2025. [Infinisst: Simultaneous translation of unbounded speech with large language model](#). *Preprint*, arXiv:2503.02969.
- Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. [StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics.
- Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. [Does Simultaneous Speech Translation need Simultaneous Models?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153.
- Sara Papi, Marco Turchi, and Matteo Negri. 2023. [AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation](#). In *INTERSPEECH 2023*, pages 3974–3978.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Peter Polák and Ondřej Bojar. 2023. [Long-Form End-to-End Speech Translation via Latent Alignment Segmentation](#). *Preprint*, arXiv:2309.11384.
- 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. Association for Computational Linguistics.
- Matt Post and Marcin Junczys-Dowmunt. 2024. [Escaping the sentence-level paradigm in machine translation](#). *Preprint*, arXiv:2304.12959.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023a. [Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors. 2023b. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics, Toronto, Canada (in-person and online).
- Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors. 2024. *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*. Association for Computational Linguistics, Bangkok, Thailand (in-person and online).
- Felix Schneider and Alexander Waibel. 2020. [Towards stream translation: Adaptive computation time for simultaneous machine translation](#). In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 228–236, Online. Association for Computational Linguistics.

Sukanta Sen, Ondřej Bojar, and Barry Haddow. 2022. [Simultaneous translation for unsegmented input: A sliding window approach](#). *Preprint*, arXiv:2210.09754.

Vaibhav Srivastav, Somshubra Majumdar, Nithin Koluguri, Adel Moumen, Sanchit Gandhi, et al. 2023. Open automatic speech recognition leaderboard. [https://huggingface.co/spaces/hf-audio/open\\_asr\\_leaderboard](https://huggingface.co/spaces/hf-audio/open_asr_leaderboard).

NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nat.*, 630(8018):841–846.

Viet Anh Khoa Tran, David Thulke, Yingbo Gao, Christian Herold, and Hermann Ney. 2022. [Does joint training really help cascaded speech translation?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4480–4487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minghan Wang, Thuy-Trang Vu, Jinming Zhao, Fatemeh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2024. [Simultaneous machine translation with large language models](#). In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 89–103, Canberra, Australia. Association for Computational Linguistics.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. [Recovering document annotations for sentence-level bitext](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9876–9890, Bangkok, Thailand. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. [Pitfalls and outlooks in using COMET](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

## A Hyperparameters

Hyperparameter	Baseline	Adapted
ASR VAD		$\times$
ASR Initial Wait		1s
ASR LCP Chunk		1s
ASR Beam Size		5
MT Wait-k		3
MT RALCP $\lambda$		0.5
MT Beam Size		10
MT Attention Head Layer		6
MT Max Buffer	80 words	
MT History Remove	20 words	1 sentence

Table 3: Inference hyperparameters for the baseline and adapted models

Hyperparameter	Value
Optimizer	8bit-AdamW (Dettmers et al., 2022)
Warm up Ratio	0.06
LR Schedule	Linear
Effective Batch Size	64
Epochs	3 or until convergence
Initial Learning Rate	$2e-4$
DoRA Dropout	0
Target Modules	$Q, V$ and Vocabulary Embeddings $E$
DoRA rank config.	$r_Q = r_K = r_E = 16$
DoRA $\alpha$	32
Bias	$\times$

Table 4: DoRA hyperparameters for the trained adapted model.