A Pocket Offline Model for Simultaneous Speech Translation as CUNI Submission to IWSLT 2026

Aziz Sharipov Ortega
Charles University, MFF, ÚFAL
azdisharipov@gmail.com

Dominik Macháček
Charles University, MFF, ÚFAL
& University of Edinburgh
machacek@ufal.mff.cuni.cz

Abstract

We implement simultaneous translation capability with the offline direct speech-to-text translation model Canary, using the state-of-the-art policy AlignAtt, and submit it to IWSLT 2026 Simultaneous Speech Translation Shared task for Czech to English and English to German and Italian.

The strengths of our system are: (1) high translation quality, outperforming similarly sized baselines both in low- and high-latency regimes in computationally unaware simulations; (2) low computational requirements, as the model has only 1B parameters; (3) multilinguality – support of 25 source and 25 target languages.

1 Introduction

In this paper, we describe a submission of Charles University (CUNI) to the IWSLT 2026 (Adelani et al., 2026) Simultaneous Speech Translation Task (Anastasopoulos et al., 2026). Our system is built on top of Canary-1B-v2[1] (Sekoyan et al., 2025) with the state-of-the-art AlignAtt (Papi et al., 2023) simultaneous policy. Following SimulStreaming (Macháček and Polák, 2025), the top-performing system in IWSLT 2025 (Abdulmumin et al., 2025) that used the same approach with Whisper model (Radford et al., 2023), we also use Silero Voice Activity Detection (VAD, Team, 2024) to reduce noise and hallucination. We validate the latency of our systems in computationally unaware simulation, and also submit a version for computationally aware simulation.

One of the strengths of our system is high multilinguality. The Canary model supports 25 source and 25 target languages for direct translation, which overlaps with three language pairs of IWSLT 2026 that we focus on: Czech to English, English to German, and English to Italian. Additionally, the model has only 1B parameters, making it a strong candidate for pocket-device deployment. Recent work has shown that ASR models of similar scale can be further quantized with minimal accuracy loss on CPU only inference (Banfic et al., 2026), suggesting the possibility of running the full simultaneous translation pipeline on constrained devices such as smartphones, eliminating a need in an external API and reducing the overall latency.

We follow the approach of repurposing offline speech translation models for simultaneous mode because it has been shown as a simple, effective, and well-performing strategy, e.g. in Papi et al. (2022) and in Macháček and Polák (2025). Offline models often offer high quality, multilinguality, and strong general robustness, which frequently outweigh their limitations such as hallucination on partial source sentences, lack of left-only context, or repeated encoding of the whole source buffer with every new incoming chunk. These limitations can be mitigated by dedicated simultaneous methods, but such methods typically require computationally expensive training, e.g. for stability on source prefixes (Niehues et al., 2018), for learning to wait or translate (Koshkin et al., 2026), or for building specific simultaneous architectures (Arivazhagan et al., 2019; Labiausse et al., 2025). We observe a higher demand for offline speech translation than for simultaneous. New, high-quality offline models are becoming available, yet they are rarely evaluated or deployed in simultaneous mode using the state-of-the-art methods. Canary is one such model, the only previously existing simultaneous implementation (Gaido et al., 2025) uses a sliding window approach that is outperformed by AlignAtt.

Our primary goal is therefore to pioneer Canary usage in simultaneous mode with AlignAtt and evaluate it against state-of-the-art systems, to test whether it can serve as a practical system and as a

[1]Throughout this paper, we use a simple term *Canary* to refer to the *Canary-1B-v2* model, although technically Canary is a family of models.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 197–203

strong baseline for future research.

We conclude we have reached this goal. The results show competitive and in some cases even superior performance to state-of-the-art baselines while being much smaller in size. Our results show improvements by over 4-5 BLEU points over the organizers' baseline on English to German and Italian and 5-8 BLEU points for Czech to English over the IWSLT 2025 best performing system (Macháček and Polák, 2025). Our implementation is integrated into the SimulStreaming project: `github.com/ufal/SimulStreaming`.

The paper is structured as follows: in the Section 2, we introduce the core model and policies and provide a relevant background. In Section 3, we dive into implementation details of our submission, while Section 4 serves as a detailed description of the frameworks and evaluation metrics used, which we then follow by the evaluation result in Section 5. Finally, we wrap up the paper with a conclusion and limitations regarding the implementation.

2 Background

Canary-1B-v2 (Sekoyan et al., 2025) is a strong multitask speech transcription and translation model with state-of-the-art performance. It outperforms *whisper-large-v3* (Radford et al., 2023) and *seamless-m4t-medium* (Communication et al., 2023) in automatic speech recognition (ASR) and automatic speech translation (AST) tasks, and on some domains and language directions is even superior to s*eamless-m4t-v2-large*, which is twice as large in size. Additionally, it supports injecting context in the decoder prompt to bias the prediction with in-domain context, and provides word-level timestamps. Unlike Whisper, Canary does not need to work with fixed size audio inputs and was instead trained to support audios in 0 to 40 seconds range, which makes it a perfect candidate for simultaneous adaptation.

AlignAtt (Papi et al., 2023) is a simultaneous policy, a method to process an offline AST or ASR model on output that is incrementally growing, one chunk at a time. The core idea AlignAtt is to use the cross-attention of the decoder to omit the suffix of hypothesis after the first token that attends to a predefined frame threshold. It has been shown to work well in long-form translation (Papi et al., 2024), but has not been previously applied on Canary.

Sliding window (Sen et al., 2022) is another simultaneous policy that we use as a contrastive baseline. Its core idea is to re-translate a window of audio input, use longest common subsequence with the previous translation to avoid repeating, and slide the window with newly available audio chunk.

3 Implementation

Let us describe the details of the implementation of our systems and the end-to-end long-form speech-to-text processing pipeline.

Simultaneous frameworks Evaluating simultaneous speech translation typically relies on frameworks that simulate real-time conditions on pre-recorded audio, enabling reproducible benchmarking. Two such frameworks support speech-to-text simultaneous systems and allow integration of new models: SimulStreaming (Macháček and Polák, 2025) and Simulstream (Gaido et al., 2025). We use primarily SimulStreaming because it provides a robust implementation of SileroVAD (Team, 2024), a streaming voice activity detection that filters non-voice parts of the input audio and allows to save computational power on processing empty input, while also avoiding potential hallucinations.

Since IWSLT 2026 task required Simulstream for computational aware evaluation, we finally transfer our Canary implementation to Simulstream to enable it. However, in case of differences, we consider our SimulStreaming implementation as the primary.

Adapting Nemo Applying the AlignAtt policy to Canary requires decoder forced prefix injection[2] for incremental output, which the NeMo API did not natively support. Hence, we highlight our contribution to the Nemo's speech framework (Harper et al.) that allows for this core incremental concept to be used with Canary. With our contribution, the prefix can be provided as an optional initial prompt to the decoder. Additionally, we fixed the bug in the cross-attention outputs for the beam decoder strategy, which allowed us to get deterministic dimension outputs and map the cross-attention scores to their respective output tokens.

Processing loop Our audio processing pipeline follows the stages of prototypical systems de-

[2]The approach is described in this thread `https://github.com/openai/whisper/discussions/117#discussioncomment-3727051`

scribed in Papi et al. (2025a), Section 3. We keep the same numbering of steps:

1.-2. Audio acquisition and segmentation using VAD is identical to SimulStreaming. Refer to Macháček and Polák (2025), Section 3.

3. Speech buffer update. The incoming chunk, which has *MinChunkSize*[3] seconds if the end of voice is not detected, or less otherwise, is concatenated with the speech buffer.

4. Hypothesis generation. Canary is an attention based encoder-decoder (AED) model. When audio is inputted, it is first encoded by the encoder. Then an initial prompt is passed to the decoder, which contains information about the source and target languages, decoder forced-prefix if any is in the buffer, as well as some other special tokens that affect final output, but are not relevant in the context of implementation, for example, whether to use timestamps. Then the model decodes the target as long as the AlignAtt policy allows. If any part of a word is inside the AlignAtt's *Frames* threshold, the word is removed from the output. If the current chunk is not final, the decoding continues until the most attended source frame is close to the end of the audio, which is indicated by the *Frames* parameter. In case the current chunk is final, we do not apply AlignAtt and output the whole generated sequence.

5. Audio and context buffers. We use the following two buffers: (1) source audio buffer, which accumulates latest 30 seconds of audio in raw form. We note that although storing the mel-features instead of raw form is more efficient, as it skips the additional preprocessing step on every new chunk, we implement a simpler buffer as we were more focused on the computationally unaware scenario, and (2) forced decoding target buffer that contains the stable part of the hypothesis that was decoded from current audio buffer.

   If the audio buffer has length of 30 seconds or more, we remove the first speech chunk from the source audio buffer. At the same time, we discard the text that was decoded with the first chunk from forced decoding. After shifting the buffer, it may happen that the audio is not entirely parallel to the forced decoded target buffer, but the results show that the model performs well with that. We leave possible improvement to further work.

[3]We mark system parameters with italics.

4 Development

Dev sets For English-German and English-Italian directions, we use the MCIF (Papi et al., 2025b) dataset, as provided by the IWSLT 2026 organizers. For Czech-to-English direction, we use the IWSLT 2026 dev set, which cosists of meetings of the Czech Chamber of Deputies (Kopp et al., 2021).

MT metrics In our initial development, we relied on BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). However, to select the final system candidates, we used COMET-XL (Guerreiro et al., 2024) as it is the top-performing and primary metric at IWSLT 2026.

Latency metrics We use computationally unaware LongYAAL (Polák et al., 2025) to select best hyper-parameters in the low- and high-latency ranges (2 and 4 seconds respectively).

5 Results

We evaluate our systems on the MCIF dev set for English-to-German and English-to-Italian, and on the IWSLT 2026 dev set for Czech-to-English. We compare against the following baselines: (1) the organizers' cascade baseline, which consists of a local agreement ASR component *Qwen3-ASR-1.7B* (Shi et al., 2026) followed by a neural MT model *Qwen3-4B-Instruct-2507* (Yang et al., 2025); (2) the Canary sliding window system as implemented in Simulstream, (3) Canary in offline mode, using the default `transcribe` function in the Nemo toolkit; and finally (4) SimulStreaming with the direct Whisper model for Czech to English direction.

Metric scores and comparison Table 1 reports our main results for the English-to-German and English-to-Italian directions. Our Canary with AlignAtt system outperforms the organizers' baseline across all four configurations – high and low latency for both language pairs – on BLEU, chrF, and XCOMET-XL. The improvements are most pronounced in the high-latency regime, where we gain over 4 BLEU points on English-to-German and more than 6 BLEU points on English-to-Italian. XCOMET-XL scores follow the same trend, with

		Reg.	BLEU	chrF	XCOMET-XL	LongYAAL (ms)
En→De	Canary ours	high	**31.73**	**60.83**	**0.8776**	3761
	baseline organizers (ctx)	high	27.66	59.92	0.8428	**3353**
	baseline organizers (no ctx)	high	27.44	59.66	0.8351	3431
	Canary offline	–	25.01	55.52	0.7932	–
	Canary sliding window	high	23.65	58.53	0.7922	2925
	Canary ours	low	20.70	52.60	**0.7744**	**1677**
	baseline organizers (ctx)	low	**22.59**	**57.51**	0.7651	1747
En→It	Canary ours	high	**43.56**	**68.32**	**0.8227**	3281
	baseline organizers (ctx)	high	37.76	65.77	0.7877	**3231**
	baseline organizers (no ctx)	high	37.28	65.44	0.7806	3300
	Canary sliding window	high	35.52	66.07	0.7729	2724
	Canary offline	–	36.78	62.22	0.7054	–
	Canary ours	low	**34.79**	62.21	**0.7618**	1972
	baseline organizers (ctx)	low	31.45	**63.03**	0.6960	**1735**
Cs→En	Canary ours	high	**32.01**	**59.26**	**0.8133**	3641
	SimulStreaming Whisper	high	24.20	50.36	0.6995	**3512**
	Canary ours	low	**27.78**	**56.68**	**0.7633**	1997
	SimulStreaming Whisper	low	22.11	49.55	0.6567	**1804**

Table 1: Dev set results of Canary with AlignAtt for simultaneous translation compared to baselines. Reg. indicates latency regime: high $<$ 4 seconds, low $<$ 2 seconds. LongYAAL is reported in milliseconds; lower is better. Our system is highlighted by green background. Baseline organizers entries differ by use of transcript context (ctx) or not (no ctx). Best individual metric results in each language pair and regime are bolded.

0.042 gains for both language pairs. Notably, this holds even when comparing against the strongest organizers' baseline configuration that uses transcript context – our system outperforms it on all quality metrics in the high-latency regime for both language pairs. In the low-latency regime, the quality gains are more modest on BLEU and chrF, especially for English to German direction, where the system falls behind the baseline, but XCOMET-XL consistently favors our system.

Figure 1 further illustrates the quality-latency trade-off on English-to-Italian. Our Canary system dominates the organizers' cascade baseline across the entire latency range, achieving higher chrF at comparable or lower latency.

Table 1 additionally reports the scores of the Simulstream authors' Canary sliding window implementation, using the configuration they provide as default (chunk 2s, window length 12, matching threshold 0.1). Our AlignAtt-based system substantially outperforms this re-translation approach: on English-to-German, we gain over 8 BLEU points in the high-latency regime, and on English-to-Italian, we gain over 7 BLEU points at a lower latency. This confirms that, while the sliding window policy is a practical and simple way to repurpose Canary for simultaneous mode, AlignAtt provides a clearly superior quality-latency tradeoff.

For the Czech to English direction, we compare our system against the SimulStreaming baseline, which uses AlignAtt with Whisper. As shown in Table 1, Canary with AlignAtt substantially outperforms this baseline in both latency regimes. In the high-latency setting, our system improves BLEU by nearly 8 points and achieves large gains in chrF as well. In the low-latency regime, we observe a similarly strong improvement of more than 5 BLEU points.

Finally, in Table 1 we report the results of running the model on the dev set in the offline mode. Canary's offline inference on long audios works as described in Sekoyan et al. (2025) in Section 6.4.1. We show that our approach works better than the offline mode and can be used instead if the running time is not a concern.

Grid-search We perform grid search to find the optimal *MinChunkSize* and *Frames* parameters to

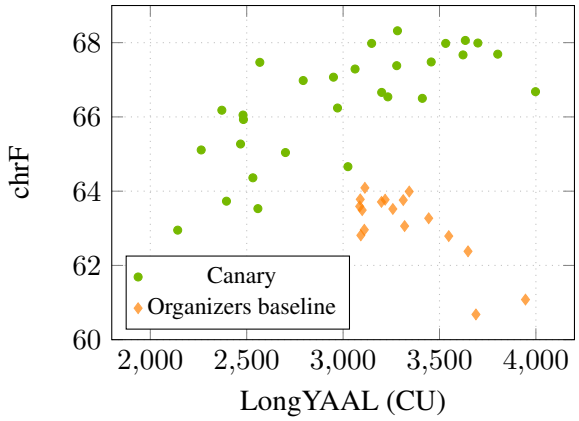

Figure 1: English-to-Italian dev chrF vs. LongYAAL (CU) for Canary vs. the organizers baseline. Each green point represents one Canary candidate from the search of *MinChunkSize* and *Frame* parameters. Orange points represent the organizers baseline system with grid-search for *segment-length* and *step-length.*

Chunk	Frame	BLEU	chrF	Latency
Latency < 2s:				
0.5	12	27.78	56.68	1997
2.5	1	23.86	53.72	1991
1.0	5	22.67	53.35	1652
1.0	4	20.13	51.85	1390
1.5	4	20.07	51.83	1899
Latency < 4s:				
2.5	20	32.01	59.14	3641
3.5	16	31.90	59.26	3921
2.0	20	31.69	59.24	3240
3.5	12	31.43	59.01	3968
3.0	12	31.36	58.95	3954

Table 2: Top-performing Canary Czech-to-English candidates for both latency regimes. We report BLEU (the higher, the better), chrF, and LongYAAL (CU) latency on the Czech-to-English dev set.

meet the latency thresholds of the IWSLT 2026 Simultaneous task, which is below LongYAAL 2000 ms for low latency, and below LongYAAL 4000ms for high latency, both in computationally unaware simulations. Table 2 showcases the influence of different combinations of the hyper-parameters on the system's performance in Czech to English.

6 Conclusion

We presented a compact and practical approach to simultaneous speech translation based on the offline Canary-1B-v2 model and the AlignAtt policy. Our main contribution is an end-to-end implementation that adapts a strong offline speech translation model to simultaneous use in the Nemo ecosystem, together with the necessary support for forced-prefix decoding and cross-attention-based truncation.

The results show that this repurposing approach is effective in practice. Across the evaluated language pairs, Canary with AlignAtt achieves competitive quality and in several settings clearly outperforms the organizers' baselines and the previously available Canary sliding-window implementation. At the same time, the system remains lightweight, with only 1B parameters, and supports a relatively broad multilingual setup, which makes it attractive for deployment in resource-constrained scenarios. Our results therefore also serve as a quality reference for what such a lightweight deployment could achieve compared to larger server-side systems and cascade pipelines. We leave the evaluation of quantized Canary in simultaneous mode as a natural next step toward practical edge deployment.

Limitations

We also report that during our experiments with the decoder prompt and context, we have not found an optimal setup to bias in-domain predictions. Injecting both forced prefix and context makes the model stall and not produce any output. We presume it could be out of training data.

Acknowledgements

This work was supported by Czech Operational Program OP JAK, the MSCA CZ project MSCA Fellowships – UK 4, CZ.02.01.01/00/22_010/0013392, "LCT".

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni,

Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Antonios Anastasopoulos and 1 others. 2026. Speech translation and metrics in 2026: Findings of the IWSLT campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Nenad Banfic, David Fan, Kunal Vaishnavi, Sam Kemp, Sunghoon Choi, Rui Ren, Sayan Shaw, and Meng Tang. 2026. Pushing the limits of on-device streaming asr: A compact, high-accuracy english model for low-latency inference.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. 2025. Simulstream: Open-source toolkit for evaluation and demonstration of streaming speech-to-text translation systems. *Preprint*, arXiv:2512.17648.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. NeMo: a toolkit for Conversational AI and Large Language Models.

Matyáš Kopp, Vladislav Stankov, Jan Oldřich Krůza, Pavel Straňák, and Ondřej Bojar. 2021. Parczech 3.0: A large czech speech corpus with rich metadata. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 293–304, Berlin, Heidelberg. Springer-Verlag.

Roman Koshkin, Je Haesung, Lianbo Liu, Hao Shi, Meng Zhao, Yusuke Fujita, and Yui Sudo. 2026. Streaming translation and transcription through speech-to-text causal alignment.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Patrick Pérez, Alexandre Défossez, and Neil Zeghidour. 2025. High-fidelity simultaneous speech-to-speech translation. *Preprint*, arXiv:2502.03382.

Dominik Macháček and Peter Polák. 2025. Simultaneous translation with offline speech and LLM models in CUNI submission to IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 389–398, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Jan Niehues and 1 others. 2018. Low-Latency Neural Speech Translation. In *Proc. Interspeech 2018*.

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)*, Bangkok, Thailand.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sara Papi, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025a. How "real" is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics*, 13:281–313.

Sara Papi, Marco Turchi, and Matteo Negri. 2023. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. *Preprint*, arXiv:2305.11408.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025b. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *Preprint*, arXiv:2507.19634.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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, Sara Papi, Luisa Bentivogli, and Ondřej Bojar. 2025. Better late than never: Meta-evaluation of latency metrics for simultaneous speech-to-text translation. *Preprint*, arXiv:2509.17349.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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*, pages 28492–28518. PMLR.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

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

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *Preprint*, arXiv:2601.21337.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.