CMU's IWSLT 2025 Simultaneous Speech Translation System

Siqi Ouyang   Xi Xu   Lei Li

Language Technologies Institute, Carnegie Mellon University, USA

{siqiouya,xixu}@andrew.cmu.edu

Abstract

This paper presents CMU's submission to the IWSLT 2025 Simultaneous Speech Translation (SST) task for translating unsegmented English speech into Chinese and German text in a streaming manner. Our end-to-end speech-to-text system integrates a chunkwise causal Wav2Vec 2.0 speech encoder, an adapter, and the Qwen2.5-7B-Instruct as the decoder. We use a two-stage simultaneous training procedure on robust speech segments curated from LibriSpeech, CommonVoice, and VoxPopuli datasets, utilizing standard cross-entropy loss. Our model supports adjustable latency through a configurable latency multiplier. Experimental results demonstrate that our system achieves 44.3 BLEU for English-to-Chinese and 25.1 BLEU for English-to-German translations on the ACL60/60 development set, with computation-aware latencies of 2.7 seconds and 2.3 seconds, and theoretical latencies of 2.2 and 1.7 seconds, respectively.

1 Introduction

CMU's submission to the IWSLT 2025 Simultaneous Speech-to-Text Translation track (Abdulmumin et al., 2025)[1] is an end-to-end model that effectively translates unbounded English speech input into German and Chinese text without speech segmentation.

Translating unbounded speech presents unique challenges. Unlike segmented speech translation, it requires the model to maintain and process the speech and translation history so that translation quality, theoretical latency, and computation cost can be balanced (Papi et al., 2024a,b). Large language models (LLMs) have recently shown strong performance in improving speech translation quality (Zhang et al., 2023; Chen et al., 2024; Huang et al., 2023; Xu et al., 2024; Ahmad et al., 2024), and modern LLMs now support

long-context inference due to architectural and algorithmic advances (Han et al., 2024; Su et al., 2024). These two advantages were recently unified in InfiniSST (Ouyang et al., 2025), which frames simultaneous translation as a multiturn dialogue and enables inference on unbounded speech with minimal computational overhead.

Our system is built upon the InfiniSST framework and consists of:

1. A chunkwise causal Wav2Vec 2.0 Large encoder (Baevski et al., 2020b), which incrementally processes the unbounded speech input.

2. The Qwen2.5-7B-Instruct LLM (Qwen et al., 2025), which receives the encoded speech features and performs simultaneous translation using a specially designed key-value (KV) cache management strategy.

However, a major limitation in speech translation research is the scarcity of high-quality parallel speech-text data. Only several hundred hours are available on resources such as EuroParl-ST (Iranzo-Sánchez et al., 2020) and CoVoST2 (Wang et al., 2021b). To scale InfiniSST training beyond this constraint, we synthesize training data by translating transcripts from automatic speech recognition (ASR) datasets into target-language text using an LLM.

Our experiments on the ACL60/60 development set (Salesky et al., 2023) demonstrate that increasing the amount of synthesized speech translation data consistently improves translation quality, with gains observed even beyond 3,000 hours of training data. Additionally, we find that Qwen2.5-7B-Instruct significantly outperforms Llama3.1-8B-Instruct (Grattafiori et al., 2024) on English-to-Chinese translation and achieves comparable performance on English-to-German translation.
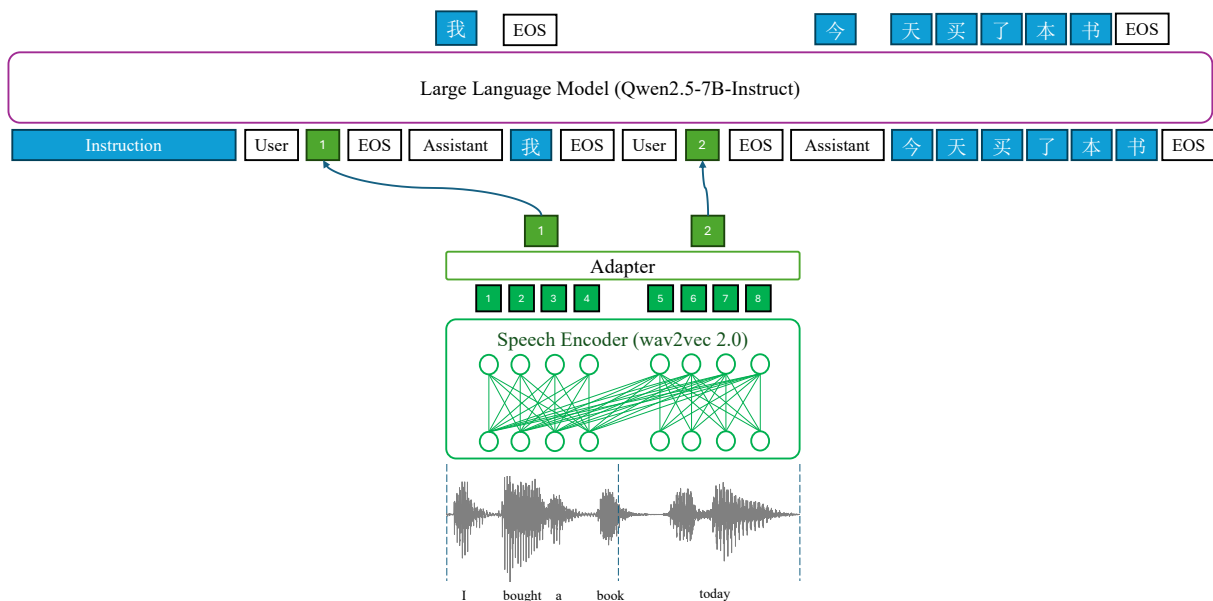
---

[1] https://iwslt.org/2025/simultaneous

Figure 1: Model Architecture

2 Task Description

The IWSLT 2025 Simultaneous Speech-to-Text Translation track[2] focuses on translating unsegmented speech into target-language text using pretrained large language models (LLMs) and speech encoders. The evaluation data consists of unsegmented ACL talks. For English-to-German, systems are additionally tested on accented speech, while a dedicated development set is provided for Czech-to-English.

Systems are evaluated on both translation quality and latency. Latency is measured using StreamLAAL (Papi et al., 2024a), while translation quality is assessed using BLEU (Papineni et al., 2002) and neural metrics such as COMET (Guerreiro et al., 2024), BLEURT (Sellam et al., 2020), etc.

We participate in two language directions: English-to-Chinese and English-to-German. For both directions, we submit models operating in the low-latency regime—achieving StreamLAAL ≤ 2 seconds for German and ≤ 2.5 seconds for Chinese on the development set.

3 System Description

3.1 Model Architecture

Our model architecture builds upon InfiniSST (Ouyang et al., 2025), a simultaneous speech translation system designed to efficiently handle unbounded streaming speech input and

generate target text incrementally. The architecture comprises three primary components: 1) a streaming speech encoder that incrementally computes representations from partial speech without redundant computations; 2) a speech-to-token embedding adapter that aligns speech representations with the LLM's token embedding space; and 3) a multi-turn LLM decoder that dynamically processes speech inputs and produces translations interactively, as shown in Figure 1.

Streaming Speech Encoder We adapt the pretrained Wav2Vec 2.0 speech encoder (Baevski et al., 2020a)[3] with several modifications. First, we replace the convolutional positional embedding with rotary positional embeddings (RoPE), due to its better performance on long sequence tasks. Second, we replace the original bidirectional attention with chunk-wise causal attention, where each chunk consists of 48 frames from wav2vec (equivalent to 960 ms). In chunk-wise causal attention, each frame can attend to frames within the same chunk and all preceding chunks, but not future ones. Third, to limit the context length and computational load, we use a sliding window approach, allowing chunk i to attend only to the hidden states from chunks within the window $[i - w^s + 1, i]$, where $w^s = 10$ represents the window size.

Speech-to-Token Embedding Adapter Outputs from the speech encoder typically have longer se-

quence lengths compared to the corresponding transcripts, and their embedding dimensions differ from those expected by the LLM. To address this, we incorporate two 1D convolutional layers with kernel size 2 and stride 2, effectively reducing the length of the encoder output sequence. Subsequently, a linear projection layer maps these convolutional outputs to match the LLM embedding space. This adapter downsamples input sequences by a factor of 4, converting each speech chunk of 48 frames into 12 embedding vectors.

Multi-turn LLM Decoder   The decoder generates the target text and emits a special EOS token when additional speech input is needed. We utilize the Qwen2.5-7B-Instruct (Qwen et al., 2025)[4] and structure the inputs using a multi-turn dialogue format. We also report results obtained with Llama-3.1-8B-Instruct (Grattafiori et al., 2024)[5] used as the decoder.

3.2   Data Synthesis

We utilize three ASR datasets for data synthesis: LibriSpeech-v12 (Panayotov et al., 2015), CommonVoice-v11.0 (Ardila et al., 2020), and VoxPopuli (Wang et al., 2021a). The English transcripts are translated into Chinese and German using the 4-bit quantized `Qwen2.5-32B-Instruct` model[6]. For LibriSpeech and VoxPopuli, whose utterances are segmented from longer speech recordings, we condition the translation on up to three preceding utterances to provide additional context. The prompt is shown below.

```
<|im_start|>system
You are a professional translator.
<|im_end|>
<|im_start|>user
Given an English sentence along with its
preceding sentences, translate the given
sentence into Chinese. Do not include any
other text.

|Preceding Sentences|
{}
|End of Preceding Sentences|

|Sentence to Translate|
```

---

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6]https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-AWQ

Dataset	# Robust Segments	Hours
LibriSpeech	174112	1393
VoxPopuli	85874	687
CommonVoice	221717	1774
Total	481703	3854

Table 1: Statistics of synthesized data for model training.

```
{}
|End of Sentence to Translate|
<|im_end|>
<|im_start|>assistant
```

Given (speech, transcript, translation) triplets, we first use Montreal Forced Aligner (McAuliffe et al., 2017) to align speech and transcript words, and then align transcript words with translation words using SimAlign (Jalili Sabet et al., 2020) with LaBSE model (Feng et al., 2022). In this way, we obtain a mapping between speech and each translation word.

Let m_i denote the right boundary timestamp of the speech segment aligned with the i-th translation word. To ensure monotonic alignment, we enforce $m_i = \max(m_i, m_{i-1})$. We then divide each utterance into fixed-duration chunks of 960 ms and construct a translation trajectory $(s_1, y_1), (s_2, y_2), \cdots$, where each s_j is a 960 ms speech chunk and $y_j = (y_{l_j}, \cdots, y_{r_j})$ is the translation span such that $m_i \leq 960 \cdot j$ for all $i \in [l_j, r_j]$.

While segmented utterances mostly consist of clean human speech, real-world scenarios often include non-speech segments. To improve model robustness, we create robust segments by slicing unsegmented speech from LibriSpeech and VoxPopuli into 30-chunk segments[7]. If a segment starts in the middle of an utterance, we shift the start to align with the utterance boundary. The trajectory for a robust segment is constructed by concatenating the trajectories of all included utterances.

Since CommonVoice consists of short, single-sentence utterances not derived from long speech, we simulate robust segments by concatenating randomly sampled utterances interleaved with randomly inserted silence intervals.

The data statistics are shown in Table 1.

---

[7]$30 \cdot 960$ ms = 28.8 seconds.

LLM	Data	English-Chinese	English-German
Llama-3.1-8B-Instruct	LS+CV	39.3 / 2092 / 2691	21.1 / 1430 / 2183
	LS+CV+VP	40.8 / 2159 / 2673	23.7 / 1503 / 2109
Qwen2.5-7B-Instruct	LS+CV+VP	44.3 / 2189 / 2739	25.1 / 1689 / 2306

Table 2: Translation quality and latency across different combinations of LLMs and training data evaluated on ACL60/60 development set. LS, CV, and VP refer to the LibriSpeech, CommonVoice, and VoxPopuli datasets, respectively. Metrics are reported as A / B / C, where A is BLEU, B is StreamLAAL, and C is StreamLAAL_CA. Incorporating synthetic speech translation data from VP leads to an improvement of at least 1 BLEU point. Additionally, Qwen2.5-7B-Instruct significantly outperforms Llama-3.1-8B-Instruct in Chinese translation, with a gain of approximately 4 BLEU points.

3.3 Training

We train our model using standard cross-entropy loss on the target translation tokens, including the special EOS token, derived from robust speech segments we constructed. Additionally, for each robust segment, we randomly sample a latency multiplier $m \leq 12$ and merge every m consecutive chunks as the data augmentation.

The training is conducted in two stages. Initially, we freeze the LLM and only train the speech encoder and adapter components. In the subsequent stage, we freeze the speech encoder, adapter and LLM, and conduct LORA finetuning (Hu et al., 2022).

3.4 Inference on Unbounded Speech

During inference, we segment the continuous input speech into fixed-length chunks of 960 ms. To manage latency, we vary the latency multiplier, ensuring translations are generated only after accumulating a predefined number of chunks. At each inference step, the newly received speech chunks are processed by both the speech encoder and the LLM, where KV caching is used to avoid redundant computations.

The speech encoder first processes new chunks along with relevant cached context. The resulting speech features are then downsampled by the adapter into a reduced set of embeddings, matching the LLM's input requirements. The LLM subsequently generates translations based on these embeddings.

The decoder uses a sliding window strategy to maintain context, combining the cached representations of initial system instructions with the most recent generated tokens similar to Han et al. (2024). We concatenate the KV cache of instruction with those of the most recent 1K tokens and apply RoPE

on top of them. Then the LLM generates translations conditioned on this combined KV cache.

4 Experiments

4.1 Setup

We use the Adam optimizer (Kingma and Ba, 2015) with cosine learning rate decay and 1,000 warmup steps. Training is conducted in two stages. In Stage 1, we update only the speech encoder and adapter, using a maximum learning rate of 2e-4. In Stage 2, we freeze the speech encoder and train the LLM with LoRA (Hu et al., 2022)[8], using a maximum learning rate of 1e-4. Each stage is trained for one epoch with a maximum effective batch size of 57.6K tokens. We leverage PyTorch Lightning[9] and DeepSpeed ZeRO[10] to train the model on a node of 8 NVIDIA L40S GPUs.

During inference, we use beam search with beam size 4, repetition penalty 1.2, and ngram_no_repeat 5. We set test-time latency multiplier to 3 for English-to-Chinese and 2 for English-to-German. The results are evaluated with BLEU, StreamLAAL and StreamLAAL_CA.

4.2 Results

Results are presented in Table 2. While the synthetic speech translation data from LibriSpeech and CommonVoice already includes over 3K hours of speech, adding additional synthetic data from VoxPopuli consistently improves BLEU scores by at least 1 point. Moreover, replacing Llama-3.1-8B-Instruct with Qwen2.5-7B-Instruct leads to a notable gain in translation quality, particularly for

---

[8]rank = 32, alpha = 16, dropout = 0.1, applied to all linear layers.
[9]https://github.com/Lightning-AI/pytorch-lightning
[10]https://github.com/deepspeedai/DeepSpeed

English–Chinese, with an improvement of over 3 BLEU points.

5 Conclusion

In this paper, we presented CMU's simultaneous speech translation system built upon the InfiniSST framework for the IWSLT 2025 SST task. Our end-to-end model employs a chunkwise causal Wav2Vec 2.0 encoder, a adapter, and the Qwen2.5-7B-Instruct decoder. We demonstrated that synthesizing training data by translating large-scale ASR datasets significantly alleviates the limitations posed by limited parallel data, achieving substantial improvements in translation quality. Our experiments indicated that the addition of synthesized data from VoxPopuli provided consistent gains, and the Qwen2.5-7B-Instruct decoder notably outperformed alternatives, particularly in English-to-Chinese translation. The proposed model effectively balances translation quality and computational latency, showcasing strong performance in a realistic, unbounded speech scenario.

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, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. LM-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008, Mexico City, Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *Preprint*, arXiv:2312.13585.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pages 498–502.

Siqi Ouyang, Xi Xu, and Lei Li. 2025. Infinisst: Simultaneous translation of unbounded speech with large language model. *Preprint*, arXiv:2503.02969.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024a. 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, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2024b. How "real" is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics*, 13:281–313.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. 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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech translation. In *Interspeech 2021*, pages 2247–2251.

Xi Xu, Siqi Ouyang, Brian Yan, Patrick Fernandes, William Chen, Lei Li, Graham Neubig, and Shinji Watanabe. 2024. CMU's IWSLT 2024 simultaneous speech translation system. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 154–159, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Xiaolin Jiao. 2023. Tuning large language model for end-to-end speech translation. *Preprint*, arXiv:2310.02050.