CMU's IWSLT 2024 Simultaneous Speech Translation System

Xi Xu[*] **Siqi Ouyang**[*] **Brian Yan   Patrick Fernandes**
William Chen   Lei Li   Graham Neubig   Shinji Watanabe
Language Technologies Institute, Carnegie Mellon University, USA
{xixu, siqiouya}@cs.cmu.edu

Abstract

This paper describes CMU's submission to the IWSLT 2024 Simultaneous Speech Translation (SST) task for translating English speech to German text in a streaming manner. Our end-to-end speech-to-text (ST) system integrates the WavLM speech encoder, a modality adapter, and the Llama2-7B-Base model as the decoder. We employ a two-stage training approach: initially, we align the representations of speech and text, followed by full fine-tuning. Both stages are trained on MuST-c v2 data with cross-entropy loss. We adapt our offline ST model for SST using a simple fixed hold-n policy. Experiments show that our model obtains an offline BLEU score of 31.1 and a BLEU score of 29.5 under 2 seconds latency on the MuST-C-v2 tst-COMMON.

1   Introduction

This paper presents CMU's submission to the IWSLT 2024 (Carpuat et al., 2024) Simultaneous Speech Translation (SST) task, focusing on streaming English speech to German text translation. Recent advancements in large language models (LLMs) have demonstrated their potential to be a strong backbone for offline ST (Huang et al., 2023; Zhang et al., 2023). In this year's submission, we build an end-to-end offline ST model with WavLM (Chen et al., 2022) and Llama2-7B-Base (Touvron et al., 2023) following the practice of LST (Zhang et al., 2023). Then we adapt the offline model for simultaneous translation.

We prepare our end-to-end ST model in the following steps:

1. Offline ST with WavLM and Llama2-7B-base.

2. Online adaptation of offline model via hold-n policy and incremental beam search.

---
[*]Equal contribution.

2   Task Description

The IWSLT 2024 SST track[1] English-German direction is a shared task for streaming speech-to-text translation of English TED talks. The task requires the system to generate the translation without modifying its previous outputs. The average lagging (AL) (Ma et al., 2019) of SST systems must be below 2 seconds on MuST-C v2.0 tst-COMMON set (Di Gangi et al., 2019). Note that AL has been modified from its original definition (Ma et al., 2020a).

Following the constraint of data and pretrained weights, we use MuST-C v2.0 as the only training set and leverage pretrained models of WavLM and Llama2-7B-Base.

3   System Description

As shown in Figure 1, our offline ST models consists of three primary components: a speech encoder, an adapter, and a LLM decoder.

For the speech encoder, we employ the WavLM model [2], which has been pre-trained on 94,000 hours data including LibriLight (Kahn et al., 2020), VoxPopuli (Wang et al., 2021) and GigaSpeech (Chen et al., 2021). We use the output of last encoder layer as the speech representation.

The modality adapter consists of two components: a length adapter and a modality adapter. The length adapter consists of two 1-dimensional convolutional layers with a kernel size of 5, a stride size of 2, padding of 2, and a hidden size of 1024. The modality adapter is a linear layer that projects the output of the length adapter to the embedding space of LLM.

We use Llama2-7B-Base as the LLM decoder. The LLM decoder takes the output of the modality adapter and autoregressively generate the target text translation.
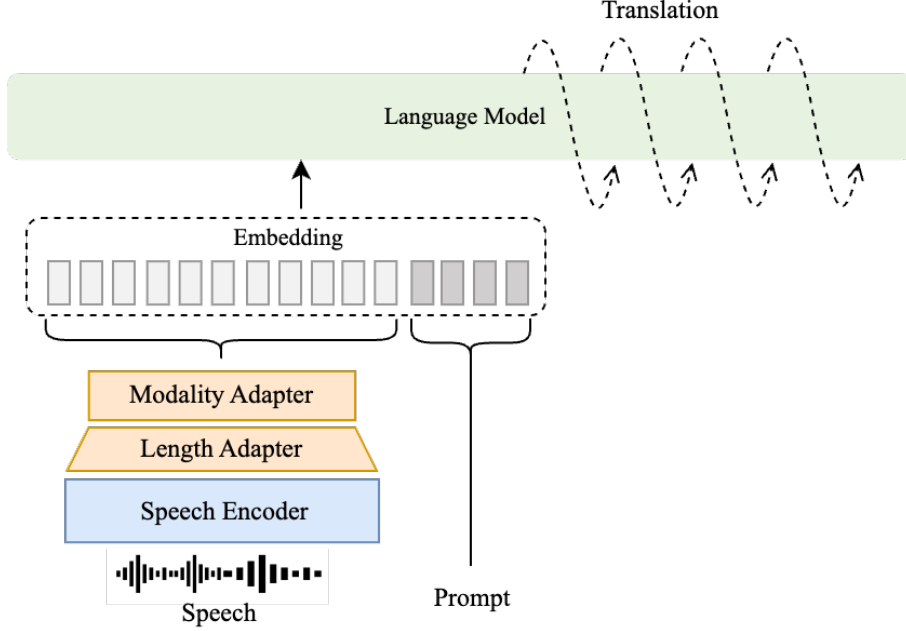
---
[1]https://iwslt.org/2024/simultaneous
[2]https://huggingface.co/microsoft/wavlm-large

Figure 1: Offline ST model architecture based on WavLM encoder and Llama2 7B decoder.

3.1 Offline Speech Translation (ST)

For each sample, given speech X^S, the reference translation X^T, and the prompt X^P, we initially transform the speech signal into a feature representation via the speech encoder:

$$H^S = \text{Encoder}(X^S), \qquad (1)$$

where $H^S = [h_1^S, \ldots, h_T^S]$ with T denoting the sequence length of the feature representation. To reconcile the length difference between the speech feature sequence H^S and its corresponding text, we downsample the speech with a length adapter.

To clarify further, the length adapter transforms H^S using a pair of 1-dimensional convolutional layers, which can be represented as:

$$Z^S = \text{Length adapter}(H^S; k, s, p, h), \qquad (2)$$

where k is the kernel size, s is the stride, p is the padding, and h denotes the number of convolutional filters. The reduced temporal dimension is $Z^S = [z_1^S, \ldots, z_N^S]$, where

$$N = \left\lfloor \frac{T - k + 2p}{s} \right\rfloor + 1, \qquad (3)$$

Next, a projector is applied to transform the speech features Z^S into E^S with the same dimension as the LLM input embedding. We use a single hidden layer as the projector,

$$E^S = \text{Linear}(Z^S). \qquad (4)$$

Finally, we feed the speech embedding E^S, translation embedding E^T, and prompt embedding E^P into the template to compose the final input E of LLM,

$$E^T = \text{Emb}(\text{Tokenizer}(X^T)), \qquad (5)$$

$$E^P = \text{Emb}(\text{Tokenizer}(X^P)), \qquad (6)$$

$$E = \begin{cases} \text{Template}(E^S, E^P, E^T) & \text{if training,} \\ \text{Template}(E^S, E^P, \tilde{E}^T) & \text{if inference,} \end{cases} \qquad (7)$$

where Emb is the LLM embedding layer, $\tilde{E}^T$ is the embedding of model's previously generated tokens.

The template is formatted as:

<P> USER: <S> ASSISTANT:<T>

where *<P>* represents the system prompt[3], *<S>* denotes the speech embedding, and *<T>* is the target reference or generated translation.

We finetune our offline ST model following a 2-stage strategy. In the first stage, we finetune the speech encoder together with the adapters, while keeping the LLM frozen. In the second stage, we finetune the entire model. We employ cross entropy loss in both stages. In addition, we apply rule-based filtering (Ouyang et al., 2022) of the dataset

---

[3]We use the following system prompt: "You are a large language and speech assistant. You are able to understand the speech content that the user provides, and assist the user with a variety of tasks using natural language. Follow the instructions carefully and explain your answers in detail."

MODEL	QUALITY	LATENCY		
OFFLINE SPEECH TRANSLATION (ST)	SACREBLEU ↑	AL ↓	LAAL ↓	LAAL_CA ↓
WavLM-LLaMA2 (Ours)	31.1	5.85	5.85	7.09
SIMUL SPEECH TRANSLATION (SST)	SACREBLEU ↑	AL ↓	LAAL ↓	LAAL_CA ↓
WavLM-LLaMA2-AlignAtt (Papi et al., 2023)	27.8	2.00	2.21	2.93
WavLM-LLaMA2 (Ours)	29.5	1.96	2.22	3.16

Table 1: Results of our English to German ST/SST models on MuST-C-v2 tst-COMMON. Latency for offline ST is calculated using a wait-k policy with k set to infinity.

Algorithm 1 Selective Output of Speech Chunk Hypotheses

```
 1: procedure SELECTIVEOUTPUT(hyps, n)
 2:     prunedHyps = {}
 3:     for c ∈ {1, . . . , C} do
 4:         W⁽ᶜ⁾ = hyps[c]
 5:         l = |W⁽ᶜ⁾|
 6:         if source_finished then
 7:             prunedHyps[c] = W⁽ᶜ⁾
 8:         else
 9:             n′ = min(n, l)
10:             W⁽ᶜ⁾_prefix = W⁽ᶜ⁾_{0:l−n′}
11:             if W⁽ᶜ⁾_prefix is not empty then
12:                 prunedHyps[c] = W⁽ᶜ⁾_prefix
13:             else
14:                 action = Read
15:                 break
16:             end if
17:         end if
18:     end for
19:     return prunedHyps
20: end procedure
```

to clean the unnecessary speaker names from the training set.

3.2 Simultaneous Speech Translation (SST)

We adapt our offline ST model for streaming inference using hold-n policy. Our scheme uses a fixed duration (e.g. 2 seconds) to compute the encoder representations on chunks of input speech. With each new chunk, we re-compute the encoder representations using the entire given input speech.

As shown in Algorithm 1, for each chunk c, we obtain the corresponding hypotheses $W^{(c)}$ using beam search given partial speech input. We then determine the number of tokens n' to withhold based on the minimum of the predefined value n and the length of the current chunk's hypotheses l. The prefix $W^{(c)}_{\text{prefix}}$ is obtained by selecting the tokens from index 0 to $l - n'$.

4 Experimental Setup

We use the AdamW optimizer with a cosine learning rate decay and a warmup ratio of 0.2. The learning rate commences at 2e-4 for the first training stage and is reduced to 2e-5 for the second stage. We train the first stage for 6 epochs and train the second stage for 1 epoch.

We employ an early stopping strategy with a patience of 6 epochs, evaluating every 1000 steps in Stage 1 and every 200 steps in Stage 2. The batch size is set to 128 for both stages. All models are trained on 4 Nividia A6000 GPUs with Deepspeed's ZeRO training strategy. The training times for the first and second stages are approximately 29 hours and 9 hours, respectively. We select the checkpoints with the lowest dev loss for testing.

For offline testing, we use a beam size of 4 to generate translations. In the simultaneous testing scenario, we set the start seconds to 2, indicating the initial wait time before processing speech chunks. We employ a hold-n strategy with n set to 7, meaning that the last 7 tokens of each chunk are withheld until more context is available. The beam size is set to 4, and the chunk size is set to 2500ms.

We evaluate translation quality using Sacre-BLEU (Post, 2018). We evaluate translation latency for SST with average lagging (AL) (Ma et al., 2020b) and length-adaptive average lagging (LAAL) (Papi et al., 2022) using SimulEval toolkit (Ma et al., 2020b).

5 Results

Table 1 shows the quality and latency of our SST system as measured on En-De tst-COMMON. We also include the offline ST performance of our model for reference. We implement the Alignatt policy (Papi et al., 2023) as a baseline for our model, we set start seconds to 2, speech segment size to 1000ms. We set number of frames to 20 and

	Wav2vec		WavLM	
LLM	Stage1	Stage2	Stage1	Stage2
TowerInstruct	-	-	29.64	-
Tower	29.35	30.53	30.11	31.64
LLaMA2	-	30.02	25.50	30.31

Table 2: SacreBLEU score of different Speech Encoder and LLMs, all models are trained on the original MuST-C 2.0 data without data cleaning.

use attention from all layers of the LLM decoder with greedy decoding.

From ST to SST, we observe a 5% quality degradation (31.1 to 29.5 SacreBLEU). However, this comes with significant latency improvements. The Average Lagging (AL) decreases from 5.85 to 1.96 seconds, a 66.5% reduction. The Length Adaptive Average Lagging (LAAL) improves from 5.85 to 2.22 seconds, a 62.1% decrease.

We also investigate the impacts of different LLMs and speech encoders, as shown in Table 2. We compare WavLM with a CTC fine-tuned Wav2vec 2.0 large model[4]. This Wav2vec model was pre-trained on 53.2k hours of untranscribed speech from LibriVox and fine-tuned on 960 hours of transcribed speech from Librispeech, as well as on pseudo-labels. Our results show that replacing Wav2vec with WavLM yields a significant improvement: a 1.1 BLEU score increase when using the Tower LLM (Alves et al., 2024) as the decoder, and a 0.3 BLEU score increase with LLaMA2 as the decoder. This suggests that the performance gains from a well-pretrained speech encoder are more pronounced when coupled with LLMs of higher translation capability.

Our analysis of the performance between different LLMs used as decoders shows that the Tower LLM[5], subjected to continued pre-training on a curated multilingual dataset of 20 billion high-quality tokens, exhibits a marked performance advantage over LLaMA2 in the initial stage of training. However, during the second stage, when the LLM backend is trainable, Tower quickly overfits, implying potential overlap between the MuST-C corpus and the data involved in Tower's pretraining. Tower

Instruct[6], which undergoes supervised fine-tuning (SFT) on instruction dataset for various translation-related tasks, achieves a slightly lower BLEU score compared to the base model. To mitigate overfitting during the second stage of training with Tower, a reduced learning rate of 7e-6 is used, compared to the 2e-5 learning rate applied to LLaMA2 training.

6 Conclusion

In this paper, we describe the submission of CMU's English to German simultaneous speech-to-text translation systems for the IWSLT 2024 Simultaneous track. We start by building a offline speech-to-text system which leverages self-supervised speech and text foundation models. We then adapt this offline model for streaming inference, enabling simultaneous speech-to-text translation.

Acknowledgements

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Marine Carpuat, Marcello Federico, Alex Waibel, Jan Niehues, Sebastian Stüker, Elizabeth Salesky, and Atul Kr. Ojha. 2024. Findings of the IWSLT 2024 Evaluation Campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*. Association for Computational Linguistics.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000

---

[4] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_vox_960h_pl.pt
[5] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

[6] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1

Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. https://github.com/facebookresearch/libri-light.

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.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. 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. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020b. Simuleval: An evaluation toolkit for simultaneous translation. In *Proceedings of the EMNLP*.

Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Ray Scott. 2015. Bridges: a uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, pages 1–8.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. On the impact of noises in crowd-sourced data for speech translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 92–97, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. 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.

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*, interspeech_2023. ISCA.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. 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.

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.