HITSZ's End-To-End Speech Translation Systems Combining Sequence-to-Sequence Auto Speech Recognition Model and Indic Large Language Model for IWSLT 2025 in Indic Track

Xuchen Wei, Yangxin Wu, Yaoyin Zhang, Henglyu Liu,
Kehai Chen[*], **Xuefeng Bai, Min Zhang,**
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
{2023311524,2023311526,2023313720,23s151043}@stu.hit.edu.cn,
{chenkehai,baixuefeng,zhangmin2021}@hit.edu.cn

Abstract

This paper presents HITSZ's submission for the IWSLT 2025 Indic track, focusing on speech-to-text translation (ST) for English-to-Indic and Indic-to-English language pairs. To enhance translation quality in this low-resource scenario, we propose an end-to-end system integrating the pre-trained Whisper automated speech recognition (ASR) model with Krutrim, an Indic-specialized large language model (LLM). Experimental results demonstrate that our end-to-end system achieved average BLEU scores of 28.88 for English-to-Indic directions and 27.86 for Indic-to-English directions. Furthermore, we investigated the Chain-of-Thought (CoT) method. While this method showed potential for significant translation quality improvements on successfully parsed outputs (e.g. a 13.84 BLEU increase for Tamil-to-English), we observed challenges in ensuring the model consistently adheres to the required CoT output format.

1 Introduction

Speech-to-text translation plays a vital role in overcoming language barriers in multilingual and international contexts, such as real-time translation during online meetings. Although translation systems for high-resource language pairs have achieved impressive performance, low-resource language pairs, particularly those involving Indic languages, continue to face significant challenges (Radford et al., 2023; Joshi et al., 2020).

This paper presents HITSZ's submission to the Indic Track of IWSLT 2025, covering bidirectional speech translation between English and three major Indic languages: Hindi, Bengali, and Tamil. An overview of the end-to-end system is illustrated in Figure 1.

Data scarcity poses a significant challenge for speech translation (ST) between English and Indic languages, primarily due to the low-resource nature of these language pairs and the reliance on data-driven neural models (Ahmad et al., 2024). Acknowledging this, we collected available parallel corpus from the official IWSLT data releases for effective end-to-end ST model training.

Cascade and end-to-end (E2E) systems represent two prominent paradigms in ST, each offering distinct advantages (Ney, 1999; Mathias and Byrne, 2006; Bérard et al., 2016). While cascaded systems typically achieve higher translation quality (Agarwal et al., 2023), E2E systems are favored for their lower latency and reduced modeling complexity (Ahmad et al., 2024; Xu et al., 2023). This work focuses exclusively on the end-to-end paradigm for the bidirectional speech translation task. We adopt an *unconstrained* setting and utilize state-of-the-art pre-trained models, including Whisper (Radford et al., 2023) and Krutrim (Kallappa et al., 2025), to develop E2E systems for both English-to-Indic and Indic-to-English directions. Although additional resources such as the IndicVoices (Javed et al., 2024) dataset are available, we deliberately exclude them due to concerns about potential overlap with the test set.

The remainder of this paper is structured as follows: Section 2 reviews related work on speech translation, particularly in low-resource and Indic language settings. Section 3 describes the datasets and data pre-processing. Section 4 introduces our end-to-end system. Section 5 presents the experimental settings, results, and analysis. Lastly, Section 6 concludes the paper.

2 Related Work

Recent advances in end-to-end speech translation (ST) have demonstrated the effectiveness of combining large pre-trained models with task-

[*] Corresponding author

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 405–411
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

specific adaptation (Wang et al., 2017; Bérard et al., 2018; Bansal et al., 2019; Wang et al., 2020; Alinejad and Sarkar, 2020), especially in low-resource and multilingual settings (Marie et al., 2019; Sun et al., 2020; Tsiamas et al., 2024; Li et al., 2025). Among them, several works stand out for their innovative training paradigms and architectural choices that have directly influenced our approach. These include NICT's submission to IWSLT 2024 (Dabre and Song, 2024), which leverages decoder-side fine-tuning of Whisper with pseudo-labels from IndicTrans2 (Gala et al., 2023); ZeroSwot, which introduces an encoder-centric alignment method for zero-shot ST (Tsiamas et al., 2024); and SALMONN, a multimodal framework that uses a lightweight training pipeline to adapt frozen encoders and LLMs through cross-modal instruction tuning (Tang et al., 2024). In what follows, we briefly review each of these works and highlight their relevance to our system design.

2.1 NICT's E2E ST System in IWSLT 2024

One of the most relevant works to our approach is the IWSLT 2024 submission by NICT, which developed end-to-end speech translation systems for English to Hindi, Bengali, and Tamil. A key contribution was their fine-tuning strategy for Whisper: instead of using human-annotated translations, they first fine-tuned IndicTrans2 to generate pseudo-translations from English transcripts. These synthetic targets were then used to train Whisper in a speech-to-text translation setting, effectively distilling knowledge and improving decoder performance beyond what reference translations alone could achieve.

2.2 ZeroSwot

Another influential work is ZeroSwot, which proposes a novel zero-shot end-to-end ST framework by aligning speech representations with the embedding space of a multilingual MT model. In their setup, the speech encoder is initialized from a CTC-finetuned wav2vec 2.0 model (Baevski et al., 2020) and trained using a combination of CTC loss and Optimal Transport loss (Graves et al., 2006; Peyré et al., 2019). The goal is to produce subword-level acoustic representations that match those expected by a frozen multilingual MT encoder (NLLB) (Team et al., 2022). In addition, a compression adapter (Liu et al., 2020) is introduced to map variable-length audio sequences into subword-aligned embeddings, bridging both length and representation mismatches between modalities. In contrast to NICT's decoder-focused fine-tuning, ZeroSwot emphasizes encoder-side alignment, enabling zero-shot translation without requiring any parallel ST data.

2.3 SALMONN

We also take inspiration from SALMONN, a multimodal framework that integrates Whisper and BEATs (Chen et al., 2023) encoders with a large language model (Vicuna) (Chiang et al., 2023) to enable general auditory understanding across speech, audio events, and music. Although SALMONN targets a broader set of audio-language tasks beyond ST, its modular design and training strategy are particularly relevant. SALMONN adopts a three-stage training pipeline—pre-training (Zhu et al., 2024), instruction tuning, and activation tuning—where only lightweight modules (Q-Former and LoRA adaptors (Li et al., 2023; Hu et al., 2022)) are updated while the encoders and LLM remain frozen. This design enables efficient adaptation with minimal parameter updates. Our work builds on this principle by leveraging pre-trained components and applying modular fine-tuning in a similarly efficient manner, tailored to low-resource, bidirectional speech translation between English and Indic languages.

3 Data

In this section, we present the statistics of the initial corpora and describe our methods for pre-processing the raw data.

3.1 Dataset

Direction	Train	Dev	Test	Total Speech Hours
en → bn	680.9	40.8	93.2	814.9
en → hi	680.9	40.8	93.2	814.9
en → ta	680.9	40.8	93.2	814.9
bn → en	158.0	1.0	1.3	160.3
hi → en	653.9	1.0	1.3	656.2
ta → en	478.2	1.0	2.2	481.4

Table 1: Statistics of dataset for training, development, and test sets. The abbreviations *en*, *bn*, *hi*, and *ta* stand for English, Bengali, Hindi, and Tamil, respectively.

We rely solely on the corpus provided by the organizers, with its statistics detailed above. Although we did not incorporate any supplementary data, our model remains *unconstrained* by leveraging the pre-trained Whisper ASR model, the Krutrim large language model, and adapters trained specifically for the spoken language translation task based on these two models. The audio segments corresponding to textual sentences are extracted from the original files based on the given offset and duration details. Post-segmentation, each dataset entry includes an audio clip in the source language, its transcription, and a translation in the target languages.

3.2 Pre-processing

We find that some audio clips in the English-to-Indic corpus are very long, indicating a very large consumption of GPU memory. To accelerate the fine-tuning process, we separate the data with English transcription length less than and above 400 characters, which allows us to increase the batch size during the training process.

4 Method

Our model builds upon the *Dhwani* model (Shah et al., 2025), which is trained for speech translation tasks in Indic languages and is itself derived from the SALMONN architecture. To effectively process and align multimodal audio data with textual outputs, the architecture integrates several specialized components. For speech signals, it leverages the Whisper speech encoder (WhisperSE) to extract robust linguistic representations. In parallel, non-speech audio inputs, such as environmental sounds and music, are processed using the BEATs encoder, which is optimized for general audio understanding.

These two audio streams are subsequently bridged to the language model via a Window-Level Query Transformer (Q-Former), which acts as a connection module to transform modality-specific features into a unified representation space. The transformed tokens are then passed to the Krutrim LLM, a 7-billion-parameter dense transformer model built on a multilingual foundation and optimized for Indic language tasks. Trained on a corpus of 2 trillion tokens with extensive coverage of native Indian languages, Krutrim demonstrates strong performance across multilingual benchmarks in both Indic and English, despite being relatively lightweight in terms of training compute.

To enable efficient domain-specific adaptation without retraining the entire model, Low-Rank Adaptation (LoRA) is employed during fine-tuning. This technique aligns the LLM's outputs with the semantics of the input audio, facilitating robust and adaptable performance.

5 Experiments and Results

This section details the experimental setup and presents the results for our monolingual speech translation models, trained individually for each translation direction. We follow the settings of the *Dhwani* model, employing the *Whisper-large-v2* model as the speech encoder and the *Krutrim-1-instruct* model as the text decoder branch.

5.1 English-Indic Translation

For the English-to-Indic translation task, we adopted a fine-tuning strategy where both the WhisperSE and the BEATs audio encoders were kept frozen. Training focused exclusively on the Q-Former module, which connects the speech encoder to the language model, and a LoRA adapter integrated into the LLM branch. We configured the LoRA adapter with a rank (r) of 8 and an alpha (α) of 32.

The learning rate schedule commenced with a linear warmup phase over the initial 3,000 training steps, increasing from a base rate of $1e^{-6}$ to the peak learning rate of $3e^{-5}$. Subsequently, the learning rate followed a cosine decay schedule, oscillating between the maximum rate ($3e^{-5}$) and a minimum rate ($1e^{-5}$), before finally decaying to the minimum rate of $1e^{-5}$.

Direction	Dev	Test
en → bn	30.61	27.00
en → hi	37.83	33.84
en → ta	25.97	22.81

Table 2: BLEU scores on the development and test set in English-to-Indic directions.

We initiated training using only the *short* audio segments from our dataset. This allowed for a larger batch size of 4, thereby accelerating the training process. Models were trained independently for three language pairs: English-to-Bengali, English-to-Hindi, and English-to-Tamil. For each pair, the checkpoint yielding the highest BLEU score on the development set was selected for subsequent incremental fine-tuning on the dataset containing *long* audio segments. Detailed results are presented in Table 2.

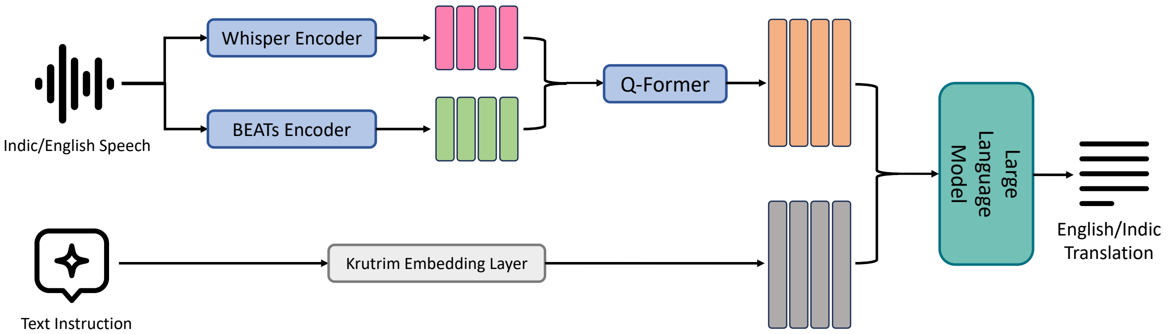

Figure 1: Overview of the our end-to-end spoken language translation system.

5.2 Indic-to-English Translation

The experimental setup for Indic-to-English translation largely mirrored the English-to-Indic configuration. However, a key difference was the absence of exceptionally *long* audio clips in the Indic-to-English corpus. Consequently, we did not employ the two-stage (*short/long*) training strategy used for the English-to-Indic directions.

Recognizing that the Whisper model exhibits comparatively lower performance on Indic languages than high-resource languages, we set the WhisperSE module to be trainable for the first epoch. Using a batch size of 1 with gradient accumulation over 4 steps helped conserve GPU memory while enabling updates to the WhisperSE, aiming for improved feature extraction from Indic audio inputs. The evaluation results are presented in Table 3.

To mitigate the challenge of limited training data and to better exploit the inherent bilingual capabilities of the LLM, we explored the Chain-of-Thought (CoT) prompting and fine-tuning technique. Specifically, this approach involved fine-tuning the model to first produce a transcription of the speech in the source language, followed by the English translation. Our findings indicate that the automatic parsing of the generated responses for the reliable extraction of the final translation output was not consistently successful.

Our experiments on the development set demonstrate that, on average, 66.54% of the responses generated by our E2E system adhere to the Chain-of-Thought (CoT) format constraints and can be successfully parsed. For the subset of responses that are parsable, results indicate notable improvements in BLEU scores. Specifically, in the Tamil-to-English translation direction, we observe a significant BLEU score improvement of 13.84 points.

Direction	Dev	Test
bn → en	25.38	25.02
hi → en	31.71	39.29
ta → en	20.93	19.27

Table 3: BLEU scores on the development and test set in Indic-to-English directions.

Direction	CoT Parsing Success Rate	BLEU Score	Δ
bn → en	68.18%	28.13	2.592
hi → en	71.00%	38.49	6.780
ta → en	60.43%	34.77	13.84

Table 4: Parsing success rate of Chain of Thought responses in Indic-to-English directions; BLEU scores of successfully parsed CoT responses on the development set; and the corresponding BLEU score improvements of the CoT method.

6 Conclusion

This paper presented HITSZ's submission to the IWSLT 2025 speech-to-text translation task in the Indic track. We leveraged recent advancements in Indic LLM by integrating the Whisper model and the Krutrim model into our end-to-end system. Future work will primarily focus on two key directions: first, enhancing the instruction-following capability of the specialized LLM for Indic languages to facilitate the development of a Spoken Language Translation system utilizing the Chain-of-Thought (CoT) method; and second, improving its generation capabilities in Indic languages to boost performance in English-to-Indic translation tasks.

Acknowledgement

We sincerely thank the shared task organizers for their efforts in designing and coordinating the task, as well as the reviewers for their valuable work that made this technical report possible. This work is supported in part by the National Natural Science Foundation of China (62276077, 62406091, and U23B2055), the Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), the Shenzhen Science and Technology Program (KQTD20240729102154066 and ZDSYS20230626091203008), and Shenzhen College Stability Support Plan (GXWD20220817123150002 and GXWD20220811170358002).

References

Milind Agarwal, Sweta Agarwal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, and 1 others. 2023. Findings of the iwslt 2023 evaluation campaign. Association for Computational Linguistics.

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.

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020.

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

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 58–68. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193. PMLR. ISSN: 2640-3498 shortConferenceName: ICML.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Raj Dabre and Haiyue Song. 2024. NICT‘s cascaded and end-to-end speech translation systems using whisper and IndicTrans2 for the Indic task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 17–22, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint*. ArXiv:2305.16307 [cs].

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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*.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K M, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, and 2 others. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6282. Association for Computational Linguistics.

Aditya Kallappa, Palash Kamble, Abhinav Ravi, Akshat Patidar, Vinayak Dhruv, Deepak Kumar, Raghav Awasthi, Arveti Manjunath, Himanshu Gupta, Shubham Agarwal, Kumar Ashish, Gautam Bhargava, and Chandra Khatri. 2025. Krutrim LLM: multilingual foundational model for over a billion people. *arXiv preprint*. ArXiv:2502.09642 [cs].

Bo Li, Shaolin Zhu, and Lijie Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167, Abu Dhabi, UAE. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR. ISSN: 2640-3498 shortConferenceName: ICML.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy. Association for Computational Linguistics.

Lambert Mathias and William Byrne. 2006. Statistical phrase-based speech translation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 517–520. IEEE.

Gabriel Peyré, Marco Cuturi, and 1 others. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR. ISSN: 2640-3498 shortConferenceName: ICML.

Sanket Shah, Kavya Ranjan Saxena, Kancharana Manideep Bharadwaj, Sharath Adavanne, and Nagaraj Adiga. 2025. IndicST: Indian multilingual translation corpus for evaluating speech large language models. In *Proc. ICASSP*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2024. Pushing the limits of zero-shot end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14245–14267, Bangkok, Thailand. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo

Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6796–6804.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.