NICT's Cascaded and End-To-End Speech Translation Systems using Whisper and IndicTrans2 for the Indic Task

Haiyue Song and Raj Dabre

National Institute of Information and Communications Technology (NICT) Hikaridai 3-5, Seika-cho, Soraku-gun, Kyoto, Japan {haiyue.song,raj.dabre}@nict.go.jp

Abstract

This paper presents the NICT's submission for the IWSLT 2024 Indic track, focusing on three speech-to-text (ST) translation directions: English to Hindi, Bengali, and Tamil. We aim to enhance translation quality in this low-resource scenario by integrating state-ofthe-art pre-trained automated speech recognition (ASR) and text-to-text machine translation (MT) models. Our cascade system incorporates a Whisper model fine-tuned for ASR and an IndicTrans2 model fine-tuned for MT. Additionally, we propose an end-to-end system that combines a Whisper model for speech-totext conversion with knowledge distilled from an IndicTrans2 MT model. We first fine-tune the IndicTrans2 model to generate pseudo data in Indic languages. This pseudo data, along with the original English speech data, is then used to fine-tune the Whisper model. Experimental results show that the cascaded system achieved a BLEU score of 51.0, outperforming the end-to-end model, which scored 19.1 BLEU. Moreover, the analysis indicates that applying knowledge distillation from the IndicTrans2 model to the end-to-end ST model improves the translation quality by about 0.7BLEU.

1 Introduction and Related Work

Speech-to-text translation is crucial for breaking the language barriers during international activities, such as translating diverse languages in online meetings. Although high-resource language pairs often achieve excellent results, the performance for low-resource language pairs remains unsatisfactory (Radford et al., 2023; Joshi et al., 2020), such as English to Indic languages. This paper presents NICT's submission to the Indic Track of IWSLT 2024, which includes translation directions from English to Hindi, Bengali, and Tamil. An overview of the cascade and end-to-end systems is illustrated in Figure 1. Data scarcity is a significant challenge for the English to Indic languages ST task due to its low-resource scenario and we are using data-driven neural models. Data augmentation on speech and text data is an efficient way to address this challenge (Shanbhogue et al., 2023; Mi et al., 2022). Assisting information such as phonetic information (Cheng et al., 2021) and spectral features (Berrebbi et al., 2022), or knowledge transferred from related languages (Anastasopoulos et al., 2022; Gow-Smith et al., 2023; Song et al., 2020) can also enhance the performance. To this end, we use data combined data from three directions rather than using them separately.

Cascade and end-to-end (E2E) systems are two popular paradigms in ST with their advantages. In general, cascaded systems show higher translation quality (Agarwal et al., 2023) and end-to-end systems usually show lower latency and less modeling burden (Xu et al., 2023). To maximize the translation quality, we adopt the cascaded way and attempt to make full use of the recent advancements in ASR and MT fields (Sperber and Paulik, 2020). Following preliminary experiments, we decided to participate in the unconstrained setting, where we leverage pre-trained models such as Whisper and Indic-Trans2 to develop our cascaded and E2E systems. Although additional datasets like IndicVoices are available for Indic languages (Javed et al., 2024), we refrain from using them due to concerns about test set overlap.

We use Whisper (Radford et al., 2023) as our ASR system. Unlike previous work (Wang et al., 2023a) who prompt Whisper without fine-tuning, we fine-tune Whisper-medium on the training data. Our results demonstrate significant improvements through fine-tuning. Although other ASR systems such as HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020) and others (Communication et al., 2023; Wang et al., 2023b) exist, we adopt Whisper for its ease of use and its ability to deliver



(a) Cascaded system training and inference process.

Training of the Speech-to-Text Translation System



Figure 1: Comparison of Cascaded and End-To-End systems.

high-quality transcriptions of English speech. We then cascade Whisper with IndicTrans2 (Gala et al., 2023) as our MT system. It supports high-quality translations across 22 popular Indic languages and outperforms the mBART50 model (Liu et al., 2020) and the M2M-100 model (Fan et al., 2020) in directions involving Indic language. Additionally, we explore the potential of the E2E system by employing knowledge distillation from the IndicTrans2 model into the Whisper model. Experimental results show that our cascaded systems are about 32 BLEU better than the E2E systems. Furthermore, E2E systems trained with distilled translations, which are obtained by translating English transcripts to Indic languages via IndicTrans2, tend to be about 0.7 BLEU points better than those using the originally provided gold standard translations.

The remainder of this paper is structured as follows: Section 2 describes the datasets and data preprocessing. Section 3 introduces our cascade and end-to-end models. Section 4 presents the experimental settings, results, and analysis. Lastly, Section 5 concludes the paper.

2 Data

We show the statistics of the original corpora and how we pre-process the raw data in this section.

Dataset 2.1

Direction	Train	Dev	Test	Total Speech Hours
$\text{en} \rightarrow \text{hi}$	44,538	7,612	7,044	95.70
$en \to bn$	5,138	1,344	1,170	16.44
$en \to ta$	7,950	2,139	2,194	22.15

Table 1: Statistics of the datasets showing the number of sentences in the training, development, and test sets alongside total speech hours.

We use only the corpus provided on the official site, with statistics shown in Table 1. We do not leverage any extra data, although our systems are built under the *unconstrained* condition. In the table, *en*, *hi*, *bn*, and *ta* represent English, Hindi, Bengali, and Tamil, respectively. We obtain the audio segments of textual sentences from the original files according to the offset and duration information. After segmentation, each data sample contains an audio segment in English, its transcription, and a translation in one of the three Indic languages. We observed that the data belongs to the spoken language domain, where the sentences are shorter compared to sentences in written texts. Moreover, there is almost no punctuation in English transcriptions and Indic language translations.

2.2 Pre-processing

To fine-tune a more robust Whisper model, we combine data from three English datasets as they belong to the same distribution.

3 Method

We describe the training and inference processes of the cascaded and E2E systems.

3.1 Cascaded System

The training and inference phases of the cascaded system are shown in Figure 1a. During training, we fine-tune the Whisper model using English audio paired with its English transcription. We then fine-tune the IndicTrans2 model using the English transcriptions and their corresponding translations in the Indic language. Although the Whisper model without fine-tuning can achieve reasonable performance, we found the format mismatch problem as presented in Figure 2. It is a type of domain mismatch between spoken language and written language, where there is less punctuation in the spoken language. However, this is prevalent in the training and development datasets, so we do not bother processing this further.

Transcription	And its only 30 years old		
Whisper output w/o fine-tuning	and it's only 30 years old.		
Whisper output w/ fine-tuning	And its only 30 years old		

Figure 2: We fine-tune Whisper to address the format mismatch problem.

During inference, the English transcription generated by the fine-tuned Whisper model is input into the fine-tuned IndicTrans2 model, which then produces the final output in the Indic language.

3.2 End-to-end System

The training and inference phases of the E2E Whisper model are shown in Figure 1b. During training, we first generate pseudo translation data using the IndicTrans2 model. We then use this pseudo data, instead of the gold transcription, to fine-tune the Whisper model. The motivation is to distill knowledge from a stronger translation model. The outputs of IndicTrans2, which are in a more consistent format, are easier for the Whisper model to learn than the human-annotated transcriptions. As shown in stage 1, we fine-tune the IndicTrans2 using English transcriptions and their translations in the Indic language. In stage 2, we generate pseudo translations for all English transcriptions in the dataset. Finally, we fine-tune the Whisper model using English audio data and these pseudo translations. During inference, we solely rely on the fine-tuned Whisper model to perform E2E ST.

4 Experiments

4.1 Settings

All our models are multilingual, achieved by combining all data into a single collection and using language indicator tokens to indicate the target language, as is the common practice. For the ASR module, we used the medium architecture of Whisper (Radford et al., 2023), which showed higher performance compared to the tiny, base, and small architectures. During fine-tuning, we set the learning rate to 1e-5, batch size to 16, and epoch size to 50. We allocated 10% of the total training steps for warmup and implemented early stopping if there was no improvement in loss after 1,000 steps, with evaluations every 100 steps on the development set. For the MT part of our experiments, we used the IndicTrans2 (Gala et al., 2023) model. We fine-tuned using the scripts provided in the IndicTrans2 library¹ including data preparation² and fine-tuning³. Using our fine-tuned IndicTrans2 model, we performed standard beam search decoding with a beam of size 5.

¹https://github.com/AI4Bharat/IndicTrans2 ²https://github.com/AI4Bharat/IndicTrans2/ blob/main/prepare_data_joint_finetuning.sh

³https://github.com/AI4Bharat/IndicTrans2/ blob/main/finetune.sh

4.2 Main Results: Submitted Systems

Table 2 presents the results of our cascaded and E2E systems on the test set.

Direction	Cascaded	E2E	Δ
English→ Bengali	52.6	10.8	41.8
$English \rightarrow Hindi$	60.5	33.0	27.5
$English \rightarrow Tamil$	39.9	13.5	26.4
Average	51.0	19.1	31.9

Table 2: BLEU scores on the test set.

Direction	Cascaded	E2E
$English \rightarrow Bengali$	50.0	7.9
$English \rightarrow Hindi$	64.1	32.1
$English \rightarrow Tamil$	41.7	12.1

Table 3: BLEU scores on the first 500 sentences from the dev set.

The scores are provided by the organizers, who do not provide a comparison with other participants at the time of writing this paper. Nevertheless, it is evident that cascaded systems outperform E2E systems by a wide margin. This indicates that data scarcity is a major problem limiting E2E system development for English to Indic language speech translation. We also provide scores for the same languages on 500 development set samples in Table 3, where we can see that there are similar trends as observed for the test set.

4.3 Impact of Distillation on E2E Systems

In Table 4, we present the differences between an E2E system trained on original translations and those trained on distilled translations. It is clear that distillation, performed by translating English transcriptions into Indic language sentences used as references for E2E systems, leads to a reasonable improvement of 0.7 BLEU.

5 Conclusion

This paper presented NICT's submission to the IWSLT 2024 English to Indic speech-to-text translation task. We took advantage of the advancements in ASR and MT where we combined the Whisper model and IndicTrans2 model in our cascaded system. In our end-to-end system, we further utilize the pseudo translation data technique, also known

Direction	E2E-Dist	E2E-Orig	Δ
English $ ightarrow$ Bengali	7.9	7.6	0.3
English ightarrow Hindi	32.1	31.2	0.9
$English \rightarrow Tamil$	12.1	11.2	0.9
Average	17.4	16.7	0.7

Table 4: BLEU scores comparison of E2E systems on the first 500 sentences from the dev set. E2E-Dist represents an E2E system trained on translated (distilled) Indic languages references, whereas E2E-Orig refers to when original references are used.

as knowledge distillation, to empower the Whisper model. Future work will focus on combining Whisper with IndicTrans2 jointly to train an even stronger speech translation system.

References

- Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 1-61, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang,

and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.
- Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel Lopez-Francisco, Jonathan D. Amith, and Shinji Watanabe. 2022. Combining spectral and self-supervised features for low resource speech recognition and translation.
- Yao-Fei Cheng, Hung-Shin Lee, and Hsin-Min Wang. 2021. Allost: Low-resource speech translation without source transcription.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul 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. *Transactions on Machine Learning Research.*
- Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER

LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech* and Lang. Proc., 29:3451–3460.
- Tahir Javed, Janki Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, R Ambujavalli, M ManickamK, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, Pratyush Kumar, and Mitesh M. Khapra. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. ArXiv, abs/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*, pages 6282–6293, Online. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speechto-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205.
- 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.
- Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. 2023. Improving low resource speech translation with data augmentation and ensemble strategies. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 241–250, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita. 2020. Pretraining via leveraging assisting languages for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics: Student Research Workshop, pages 279–285, Online. Association for Computational Linguistics.

- Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.
- Minghan Wang, Yinglu Li, Jiaxin Guo, Zongyao Li, Hengchao Shang, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2023a. The HW-TSC's speechto-speech translation system for IWSLT 2023. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 277–282, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023b. Viola: Unified codec language models for speech recognition, synthesis, and translation.
- 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, IJCAI-23*, pages 6796–6804. International Joint Conferences on Artificial Intelligence Organization. Survey Track.