JU-CSE-NLP's Cascaded Speech to Text Translation Systems for IWSLT 2025 in Indic Track

Debjit Dhar[1†], **Soham Lahiri**[1†], **Tapabrata Mondal**[1], **Sivaji Bandyopadhyay**[1*]
[1]Jadavpur University, Kolkata, India
[†]Authors contributed equally
[*]**Correspondence:** sivaji.cse.ju@gmail.com

Abstract

This paper presents the submission of the Jadavpur University Computer Science and Engineering Natural Language Processing (JU-CSE-NLP) Laboratory to the International Conference on Spoken Language Translation (IWSLT) 2025 Indic track, addressing the speech-to-text translation task in both English-to-Indic (Bengali, Hindi, Tamil) and Indic-to-English directions. To tackle the challenges posed by low-resource Indian languages, we adopt a cascaded approach leveraging state-of-the-art pre-trained models. For English-to-Indic translation, we utilize OpenAI's Whisper model for Automatic Speech Recognition (ASR), followed by the Meta's No Language Left Behind (NLLB)-200-distilled-600M model finetuned for Machine Translation (MT). For the reverse direction, we employ the AI4Bharat's IndicConformer model for ASR and IndicTrans2 finetuned for MT. Our models are fine-tuned on the provided benchmark dataset to better handle the linguistic diversity and domain-specific variations inherent in the data. Evaluation results demonstrate that our cascaded systems achieve competitive performance, with notable BLEU and chrF++ scores across all language pairs. Our findings highlight the effectiveness of combining robust ASR and MT components in a cascaded pipeline, particularly for low-resource and morphologically rich Indian languages.

1 Introduction and Related Work

Speech to text translation (STT) has long been the interest of Natural Language Processing (NLP) researchers particularly due to the huge number of languages spoken worldwide. However, a major part of research has been invested on translation between European languages and some Asian languages such as Chinese. Thus, translation from English to Indic languages and vice-versa is of considerable importance. This is further highlighted by the fact that Indic languages like Hindi, Bengali and Tamil have several speakers worldwide (615, 228 and 90.8 million respectively) (Ahmad et al., 2024). The top performers (NICT (Dabre and Song, 2024) and HWTSC (Wei et al., 2024)) of IWSLT 2024 Indic Track Competition showed the importance of finetuning ASR and MT models in a cascaded system. We present in this paper, cascaded models for translation from English to Indic languages (Bengali, Hindi, Tamil) and vice-versa. We also fine-tune our Neural Machine Translation (NMT) models using IWSLT 2025 Indic Track Training Dataset so as to obtain better results as compared to the pretrained checkpoints. The visual description of our system is shown in Figure 1 . We prefer the cascaded system over the End to End (E2E) system for two reasons. Firstly, most speech translators are not trained on indic speech and hence would require lot of resources to achieve the baseline performance that we have in cascaded systems. Secondly even in the English to Indic track, it has been shown in (Ahmad et al., 2024) that cascaded models completely outperform their end to end counterparts. Transcription and translation is carried out entirely on the basis of the given segmentation timestamps in the train, dev and test sets of the IWSLT 2025 Indic Track.

2 Dataset Description

The dataset for each language pair was given in three parts for both train and development sets. These included the wav files containing the audio of the speaker, yaml files containing the audio metadata and the segmentation information and the text files containing the corresponding transcriptions and translations of the segments. The punctuations were present only in the translated text. It was observed that the source files in English to Indic direction were same for the three languages. The test datasets contained only wav files and yaml files.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 201–205

Direction	Train	Dev	Test
English -> Bengali	205209	11671	36245
English -> Hindi	205209	11671	36245
English -> Tamil	205209	11671	36245
Bengali -> English	64868	395	866
Hindi -> English	248872	397	579
Tamil -> English	211303	457	956

Table 1: Number of segments given in the yaml files

3 Methodology

We have employed cascaded systems for both English to Indic (Hindi, Bengali, Tamil) as well as Indic (Hindi, Bengali, Tamil) to English Translations. We describe the preprocessing, finetuning and inference procedure of our models in the subsequent subsections.

3.1 Pre-processing

Before using our Automatic Speech Recognition (ASR) model we perform the following preprocessing steps on the audio files. Firstly, we normalize the audio volume in [-1,1] range. This is followed by applying a biquad low pass filter for noise reduction (with cutoff at 3kHz). Thirdly, we amplify the entire audio by 10dB. This is required for those cases where the speaker's voice is inaudible or unclear. We observed a notable reduction in WER and an improvement in SacreBLEU scores as a result of acoustic pre-processing shown in Table 2. The pre-processing, inference pipeline is shown in Figure 1.

System	Score	Bengali	Hindi	Tamil
English to Indic (without Preprocessing)	WER	27.75	27.75	27.75
	SacreBLEU	16.81	17.72	11.91
English to Indic (with Preprocessing)	WER	21.09	21.09	21.09
	SacreBLEU	21.79	24.46	12.81
Indic to English (without Preprocessing)	WER	56.57	37.42	65.81
	SacreBLEU	26.05	31.19	27.78
Indic to English (with Preprocessing)	WER	48.32	36.93	56.17
	SacreBLEU	39.17	46.28	37.69

Table 2: WER and SacreBLEU scores with and without preprocessing using pretrained checkpoints

3.2 English to Indic System

For ASR we use Whisper Small model by OpenAI (Radford et al., 2023) and for NMT we use the NLLB-200-distilled-600M variant) by Meta (Costa-Jussà et al., 2022). We intentionally choose the Small version of Whisper as this model gives the best SacreBLEU score (Post, 2018) when paired with our NMT and also has comparable Word Error Rate (WER) on the given dev set as shown in Table 3. After comparison with other SOTA models such as Helsinki Opus (Tiedemann and Thottingal, 2020) , we find that NLLB-200-distilled-600M gives us the best SacreBLEU score on the dev set Table 4. However, it is observed that the NMT model does not give a reasonable accuracy while using the existing checkpoints. Hence, we resort to finetuning the NLLB-200-distilled-600M model (Xinyuan et al., 2023) on the train set as given in IWSLT 2025 competition. Firstly, finetuning is carried out only on the NMT model. Here we use the source and target texts given in the competition train dataset. As for the pipeline, we run the Whisper Small model in inference mode and use its output as input for the finetuned NLLB model thus obtaining the translated text.

3.3 Indic to English System

For the Indic-to-English system, we utilize AI4Bharat's IndicConformer (`https://github.com/AI4Bharat/IndicConformerASR`) as the ASR model. For Indic-to-English translation, we employ IndicTrans2 (Gala et al., 2023) as the NMT model. During training, we fine-tune the NMT model using Indic transcriptions and their corresponding English translations. This fine-tuning results in a noticeable improvement in SacreBLEU scores.

We arrived at this choice of ASR and NMT models after conducting extensive experiments with different combinations of ASR and NMT models. For the MT system we tried using mBART (Tang et al., 2020) and NLLB200 but observed a much lower performance as compared to IndicTrans2. For ASR selection, we chose the system with the lowest WER, while for NMT, we evaluated the SacreBLEU score of the entire cascaded system on the development set to determine the best-performing model.

4 Experiments

The fine-tuning was conducted on a multi-GPU setup using Kaggle GPU T4x2 for efficient parallel-processing. To optimize training, audio-related metadata were removed.

4.1 Settings for English-Indic System

Due to resource constraints, we are it was not possible to finetune Whisper Small ASR on English to Indic system. NLLB200 is transformer based

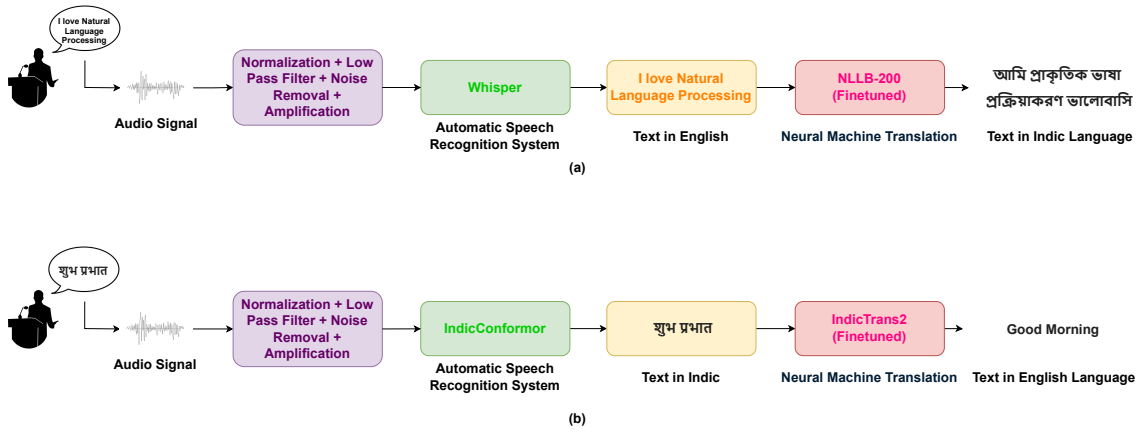

Figure 1: Overview of the proposed Multilingual Speech Translation Pipeline: (a) English-to-Indic flow using Whisper and finetuned NLLB-200; (b) Indic-to-English flow using IndicConformer and finetuned IndicTrans2.

MT model developed by Meta AI as a part of the No Language Left Behind initiative. This model had been evaluated on the Flores200 dataset and had demonstrated promising results particularly for under-represented languages. Hence, we choose this model for the machine translation part of our cascaded system. We finetune the NLLB model separately for Bengali, Hindi and Tamil target texts. Moreover in each case finetuning is done incrementally taking 20000 samples from train set each time. The finetuning was done with learning rate as 2e-5, batch size as 2, beam size 5, weight decay as 0.01 and for 5 epochs. The HuggingFace interface of NLLB was used for the finetuning procedure. Table 6 shows our final results on the dev set for English to Indic system.

4.2 Settings for Indic-English System

From Table 5, we identify the best-performing pretrained NMT model for each specific Indic-English pair and proceed to fine-tune them accordingly. Due to resource constraints, it was not possible to fine-tune the IndicConformer ASR model for the Indic-to-English system. Instead, we focused on fine-tuning the IndicTrans2 model, a state-of-the-art multilingual neural machine translation system tailored for Indic languages, using a LoRA-based parameter-efficient strategy (Hu et al., 2022; Patil et al., 2024).

IndicTrans2 is a transformer-based sequence-to-sequence model pretrained on large-scale translation corpora and supports multiple Indic languages. For fine-tuning, we use the development set of IWSLT 2025 Indic Track. Sentence-aligned parallel data was loaded and preprocessed using the IndicProcessor (Gala et al., 2023), which performs normalization and script standardization appropriate to each language. Tokenization was performed using the AutoTokenizer compatible with the base IndicTrans2 model. Preprocessing also involved truncating long sequences to adhere to the model's maximum input length.

To reduce training costs and memory usage, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022; Patil et al., 2024) via the peft library. Only a small subset of model parameters—specifically the attention projection matrices—were updated, while the rest of the model remained frozen. This allowed efficient adaptation to new data without full-scale retraining.

Fine-tuning was carried out using the Seq2SeqTrainer from the Hugging Face Transformers library. We used mixed-precision training (fp16) for computational efficiency. The model was evaluated on a held-out validation set after each epoch using automatic evaluation metrics such as BLEU (Papineni et al., 2002)and chrF++. (Popović, 2015) Early stopping was applied based on validation loss to prevent overfitting.

The learning rate was set to 2×10^{-5}, with a batch size of 4 per device, adjusted for multi-GPU training and LoRA rank and alpha of 4 and 16 respectively. The model was trained for 5 epochs, with a weight decay of 0.01 and a beam size of 5. Table 6 presents the final results on the development set for Indic-to-English translation.

5 Results

The final finetuned results on the dev set are given in Table 6. The WER observed on the dev sets for

ASR Model	Word Error Rate %
Whisper Tiny	25.63
Whisper Small	21.09
Whisper Medium	20.83
Whisper Base	23.16
Whisper Large-V2	20.74
Whisper Large-V3	20.62

Table 3: WER on the Source English text of combined English to Indic dev datasets

NMT Models	En-Bn	En-Hi	En-Ta
NLLB 200	21.79	24.46	12.81
Helsinki Opus	13.49	12.30	-

Table 4: SacreBLEU Scores on Dev Set using pretrained checkpoints (Helsinki does not have En-Ta checkpoints)

Bengali, Hindi, and Tamil using IndicConformer were 48.32%, 36.93%, and 56.17%, respectively. (Abdulmumin et al., 2025) The results on the test set as calculated by IWSLT are given in Table 7.

6 Limitation

Due to resource constraints, it was not feasible to fine-tune the ASR models to reduce the word error rate, which directly affects the quality of input provided to the NMT system. Additionally, the fine-tuning of the NMT models was limited to a maximum of five epochs, further constraining potential improvements in translation performance.

7 Conclusion and Future Work

This paper presented JU-CSE-NLP's submission to the Indic Track of IWSLT 2025. We have highlighted the detailed methodology of our preprocessing, finetuning and inference procedures in our paper which will help further research and system development in the field of speech translation of Indic languages. Our results are also quite reasonable comparing with previous year's performance in the same track (Ahmad et al., 2024) as shown in Table 8.

NMT Model	Bn-En	Hi-En	Ta-En
mBART	8.92	22.02	13.77
NLLB 200	21.09	0.50	9.29
IndicTrans2	42.80	31.19	27.78

Table 5: SacreBLEU scores on IndicConformer output of Dev Set using pretrained checkpoints

System	Bengali	Hindi	Tamil
English to Indic	44.54	39.04	38.82
Indic to English	39.17	46.28	37.69

Table 6: SacreBLEU scores of finetuned cascaded systems on Dev Set

System	Score	Bengali	Hindi	Tamil
English to Indic	BLEU	51.70	57.61	36.17
	chrF++	74.58	72.98	73.81
Indic to English	BLEU	23.69	44.13	17.66
	chrF++	53.99	67.91	49.34

Table 7: BLEU and chrF++ scores of cascaded systems on Test Set

In future work, we aim to conduct a comprehensive error analysis of our results to identify key areas for improvement and further enhance system performance. We also plan to apply knowledge distillation techniques from our cascaded systems to state-of-the-art end-to-end models, with the goal of achieving competitive performance in those frameworks. Additionally, we intend to extend our current Speech-to-Text (S2T) system into a full Speech-to-Speech Translation (S2ST) system. While our present approach is monolingual for each language pair, we aim to develop a multilingual system capable of handling multiple languages in all translation directions. Furthermore, we plan to incorporate language-specific features to improve translation quality and robustness.

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*

Team ID	En-Bn	En-Hi	En-Ta
NICT	52.63	60.54	39.84
HWTSC	35.04	47.14	30.79
NITK	4.46	19.77	11.76
Ours	**51.70**	**57.61**	**36.17**

Table 8: Comparison with BLEU score of IWSLT 2024 English - Indic Unconstrained Cascaded Systems

(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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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.

Pranamya Patil, Raghavendra Hr, Aditya Raghuwanshi, and Kushal Verma. 2024. SRIB-NMT's submission to the Indic MT shared task in WMT 2024. In *Proceedings of the Ninth Conference on Machine Translation*, pages 747–750, Miami, Florida, USA. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Bin Wei, Zongyao Li, Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Hao Yang, and Yanfei Jiang. 2024. HW-TSC's speech to text translation system for IWSLT 2024 in Indic track. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 53–56, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Henry Li Xinyuan, Neha Verma, Bismarck Bamfo Odoom, Ujvala Pradeep, Matthew Wiesner, and Sanjeev Khudanpur. 2023. JHU IWSLT 2023 multilingual speech translation system description. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 302–310, Toronto, Canada (in-person and online). Association for Computational Linguistics.