CDAC-SVNIT submission for IWSLT 2025 Indic track shared task

Mukund K Roy[1,2], Karunesh K Arora[1], Praveen Kumar Chandaliya[2], Rohit Kumar[2], Pruthwik Mishra[2],
[1]SNLP Lab, CDAC Noida, India, [2]SVNIT Surat, India,
Correspondence: mukundkumarroy@cdac.in, karunesharora@cdac.in, pkc@aid.svnit.ac.in,
rohitkumar@aid.svnit.ac.in, pruthwikmishra@aid.svnit.ac.in

Abstract

In this paper, we design a Speech-to-Text Translation (ST) system to translate English into Hindi, Bengali, and Tamil, and vice versa. We explore both cascaded and End-to-End (E2E) approaches as part of the IWSLT 2025 Indic shared task. In the cascaded systems, we leverage the pre-trained Wav2Vec2 model from AI4Bharat's Vakyansh project, and then fine-tune it for Automatic Speech Recognition (ASR). The resultant ASR outputs are then translated using the adapted IndicTrans2 Neural Machine Translation (NMT) model with IWSLT task-specific data. In the E2E approach, we train models from scratch using only the IWSLT dataset, leveraging the Fairseq Speech Translation framework which uses transformer-based encoder-decoder architecture optimized for multilingual speech inputs. In the paper, the performance of these two distinct approaches in handling low-resource Indic speech translation tasks is compared. Although in the E2E approach, the pre-trained Acoustic model is not leveraged, its results in the En-Indic setting are impressive. However, this approach does not perform well in the Indic-En setting due to lack of sufficient training data. On the other hand, the cascaded approach leverages pre-trained models and outperforms for all language pairs.

1 Introduction

In a global and borderless economy, seamless communication is essential, with speech being the most natural medium. Overcoming language barriers through intelligent systems is crucial for real-time interaction and bridging the digital divide (Arora et al., 2013). Speech-to-text translation has a vital role to play in facilitating communication across language barriers. Recent advancements in the area of speech technology have resulted in state-of-the-art performance in the speech recognition task (Baevski et al., 2020a; Radford et al., 2022) and machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2016; Vaswani et al., 2017) for almost all major languages. This encourages the advent of direct speech translation of speech, leading to the rise of two different paradigms of achieving the same. They are: *Cascaded* and *End-to-End* speech translation. In the cascaded speech-to-text (ST) translation paradigm, the task of translating speech from a source language to text in a target language is broken down into two distinct modules. The recent rise of cascaded ST systems (Mujadia and Sharma, 2023; Prakash et al., 2023; Mhaskar et al., 2023) for translating the educational content in Indian languages show the effectiveness of this approach.

Automatic Speech Recognition (ASR): The input speech in the source language is first transcribed into text using an ASR system.

Machine Translation (MT): The transcribed source language text is then translated into the target language using a Neural Machine Translation (NMT) system.

This pipeline-based approach is advantageous for modular development, allowing the ASR and MT components to be trained independently and optimized using speech datasets, even when parallel ST corpora are limited. However, a limitation of cascaded ST is the potential propagation of errors from ASR to MT, where transcription errors can negatively impact the translation quality. End-to-End Speech Translation is another paradigm that directly translates spoken utterances in one language into text in a target language, bypassing intermediate steps such as ASR and MT (Weiss et al., 2017). This approach enables the model to learn joint representations that capture both acoustic and linguistic features, resulting in efficient inference and reduced error propagation compared to traditional cascaded pipelines (Sperber and Paulik, 2020). Leveraging architectures such as encoder-decoder transformers, end-to-end ST

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 180–185

Lang-pair	Train		Dev		Test	
	# of Audios	Total duration	# of Audios	Total duration	# of Audios	Total duration
En–Hi	205,201	680h 54m	11,669	40h 47m	36,245	93h 13m
En–Bn	205,203	680h 54m	11,671	40h 48m	36,245	93h 13m
En–Ta	205,203	680h 54m	11,671	40h 48m	36,245	93h 13m
Hi–En	248,872	653h 52m	397	0h 59m	579	1h 20m
Bn–En	64,868	157h 57m	395	1h 0m	866	1h 15m
Ta–En	211,303	478h 9m	457	0h 59m	956	2h 11m

Table 1: Statistics of the speech translation dataset provided for the IWSLT 2025 Shared Task. Durations are shown in hours and minutes.

systems are trained on pairs of ST translation data, allowing them to implicitly align and map source audio to target textual content (Dong et al., 2018). This methodology has shown promising results in multilingual and low-resource settings, especially when supported by self-supervised pretraining and transfer learning from large ASR and MT models (Bérard et al., 2018; Wang et al., 2020). However it faces key limitations such as the scarcity of parallel speech-to-translation data. These models simultaneously learn acoustic processing, language understanding, and translation, leading to slower convergence and reduced performance, especially in low-resource settings. Additionally, it also struggle with varied pronunciations and code-switching due to the lack of intermediate transcripts which could be normalized or transliterated. So continuing the work towards low-resource language pairs, Inaguma et al. (2020) propose a multilingual end-to-end speech translation framework utilizing shared encoder and decoder components. This architecture leverages parameter sharing and cross-lingual transfer learning, leading to significant improvements in translation quality. Salesky et al. (2021) focus on speech translation in low-resource settings and explored strategies such as multilingual fine-tuning and data augmentation. Their findings indicate that these methods can effectively compensate for limited training data and improve translation accuracy across modalities.

2 Data

In the IWSLT 2025 Indic Speech Translation Shared Task (Abdulmumin et al., 2025), multilingual speech translation dataset spanning six distinct language pairs involving English and three indic languages are released. Table 1 shows the statistics of the audio corpus that is aimed to support both training and evaluation of speech translation systems in low-resource settings. For the training set, each language pair offers a substantial volume of audio data. The En-Bn (English–Bengali), En-Hi (English–Hindi), and En-Ta (English–Tamil) pairs have 205,000 audio samples, amounting to approximately 681 hours. The Hi-En (Hindi–English) direction includes the largest dataset, comprising 248,872 audio segments with a total duration of approximately 654 hours. The Ta-En (Tamil–English) pair includes 211,303 training audios, summing up to 478 hours, while the Bn-En (Bengali–English) dataset is slightly smaller in size with 64,868 samples and 158 hours of speech data. For validating the models, the devset is also provided for each language pair which is approximately 6% in case of En-Indic pairs, while it is below 1% in case of Indic-En pairs. For this shared task, no other synthetic data has been used by performing Machine Translation on source language ASR output or synthesizing speech from the target language text.

3 Methodology

3.1 Cascaded S2T

For this experiment, we finetune CLSRIL-23[1] [2] (Gupta et al., 2022), a self-supervised model that is designed to leverage cross-lingual speech representations from raw audio dataset. The pre-training dataset consists of approximately 10,000 hours of audio data across 23 Indic languages. The architecture of CLSRIL-23 is based on Wav2Vec 2.0 (Baevski et al., 2020b), where the base version has 12 transformer blocks with 768 dimensional feaure vector size and 12 attention heads. It comprises of multi-layer convolutional feature encoder that processes raw audio inputs into latent speech representations. These representations are then fed into a Transformer network, which captures con-

[1] https://github.com/Open-Speech-EkStep/vakyansh-models

[2] https://github.com/Open-Speech-EkStep/vakyansh-wav2vec2-experimentation

textual information over the entire sequence. The model is trained on a contrastive loss function to distinguish true quantized latent representations from distractors, facilitating the learning of robust speech representations.

ASR Fine-Tuning: For the IWSLT Indic track task, we fine-tune CLSRIL-23 using the dataset provided by the organizers, which included parallel speech and text data for the target Indic languages. The fine-tuning process involves the following steps:

- **Data Preparation**: In this step, we align and preprocess the provided speech and text pairs to ensure the compatibility with the input requirements of our model.
- **Model Finetuning**: We initialize the pre-trained CLSRIL-23 model and add a fully connected layer on top of the transformer block to perform character-level classification. During fine-tuning, we keep the weights of the feature encoder frozen, allowing only the transformer and classification head to be updated. For fine-tuning the model, the learning rate is kept at $3e^{-5}$ with a batch size of 32, and we train it for 50 epochs to optimize performance on the task.
- **Evaluation**: Using the provided validation set, the effectiveness of model are evaluated in terms of Word Error Rate (WER) and Character Error Rate (CER).

By fine-tuning on the provided dataset, the model refine its previously learned features to better capture the unique patterns and properties present in the data.

Machine Translation

For the text translation component of the speech translation pipeline, we fine-tune the IndicTrans2[3] (Gala et al., 2023), a multilingual NMT model. It is capable of translating from English to 20 Indic languages and vice-versa. The model has 1.1 billion parameters pre-trained on a mixture of parallel corpora, combining general-domain, news, and publicly available data sources, making it suitable for fine-tuning it for this shared task on Indic low resource languages. The process of fine-tuning is as the following:

- **Data Preparation**: We perform script normalization, Unicode standardization, whitespace cleanup, and special character filtering to reduce the noise present in the dataset. To enable multilingual translation to the target language, we prepend language-specific prefix tokens to the source sentences, following the original IndicTrans2 multilingual setup. Finally, we tokenize the processed data using the SentencePiece tokenizer (Kudo and Richardson, 2018) released with IndicTrans2, ensuring compatibility with its subword vocabulary and avoiding out-of-vocabulary (OOV) issues during training and inference.
- **Model Fine-tuning**: We use a deep transformer model designed to handle the complexity of multilingual neural machine translation. This architecture comprises of 18 encoder and 18 decoder layers, each with a hidden dimensionality of 1024 and Feed-Forward Network (FFN) layers of size 8192. The model is fine-tuned using a learning rate of $3e^{-5}$ and AdamW (Loshchilov and Hutter, 2017) optimizer, with a weight decay of 0.01 to prevent overfitting. We enable mixed precision training to make efficient use of GPU memory and accelerate computation. For evaluation, we monitor performance on the validation set using the SacreBLEU (Post, 2018) metric, which provide a reliable estimate of translation quality across different language pairs.

3.2 End-to-End S2T

For our end-to-end speech translation experiments, we use a small-scale transformer-based encoder-decoder model available in the Fairseq Speech-to-Text framework[4] (Ott et al., 2019; Wang et al., 2020). This model utilizes an encoder embedding dimension of 256 and a feed-forward network with a dimension of 2048. Both the encoder and decoder use 4 attention heads and a dropout rate of 0.1 for regularization. The model inherits from the base architecture, which by default configures 6 layers each for the encoder and decoder. This configuration is effective from the starting point for training and evaluating end-to-end speech translation systems, especially in low-resource or computationally constrained settings.

- **Data Preparation**: We use the provided script

[3] https://github.com/AI4Bharat/IndicTrans2/

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/speech_to_text

Cascaded-Unconstrained-monolingual			E2E-Constrained-monolingual		
Lang-pair	**ChrF++**	**BLEU**	**Lang-pair**	**ChrF++**	**BLEU**
En-Hi	64.1749	44.093	En-hi	54.4822	34.6119
En-Bn	65.2117	36.9565	En-bn	58.2243	31.5668
En-Ta	66.1503	29.341	En-ta	56.0757	21.3467
Hi-En	67.0583	41.0425	Hi-En	42.9691	15.4153
Bn-En	44.8855	14.7731	Bn-En	14.3009	0.459
Ta-En	41.1644	15.7004	Ta-En	26.2496	5.0473

Table 2: Comparison of translation performance between Cascaded-Unconstrained and End-to-End Constrained systems using ChrF++ and BLEU scores.

in the framework to prepare the speech dataset for training. It processes the audio and transcription files organized in each language pair's respective directory and splits into train and validation sets. For each audio segment, it extracts the log Mel filterbank features and generates corresponding manifest files (stored in a tab separated format) with the metadata. The script also builds a vocabulary file using SentencePiece and a config file needed for the Fairseq training.

- **Model Training**: For training the speech translation model on the dataset, we use the speech to text transformer architecture available in the Fairseq library. We set the maximum number of tokens per batch to 40000 to efficiently utilize GPU memory and the training is capped at 200 epochs. To improve generalization, we apply label smoothing with a value of 0.1 and use a dropout rate of 0.3 to regularize the model. The optimizer is Adam (Kingma and Ba, 2014), with a learning rate of $2e^{-3}$, and gradient clipping is set at 10.0 to prevent exploding gradients. For inferencing, we take the average of the last 10 checkpoints as Vaswani et al. (2017) proved that the averaged checkpoint performs better than the single best checkpoint. SacreBLEU is used for scoring the performance of the models.

4 Experimental Results

For the IWSLT 2025 Indic Speech Translation Shared Task (Abdulmumin et al., 2025), we participate in two different settings: a) Unconstrained Cascaded and b) Constrained End-to-End speech-translation track. The experiments are not multilingual, but individual language-pairs are trained separately. We conduct experiments on all six language pairs: English to Hindi (en-hi), Bengali (en-bn), Tamil (en-ta), and the reverse directions hi-en, bn-en, ta-en respectively. The results of our experiments are presented in Table 2, showing ChrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) scores for each language pair across both the cascaded and E2E settings.

English-to-Indic (en-hi, en-bn, en-ta): The cascaded system consistently outperform the E2E system across all the language pairs. For example, en-hi achieves a BLEU of 44.09 and ChrF++ of 64.17 in the cascaded setup, compared to 34.61 BLEU and 54.48 ChrF++ in the E2E setup. Similarly, the en-bn model scores 36.95 BLEU (ChrF++: 65.21) in the cascaded mode versus 31.56 BLEU (ChrF++: 58.22) in E2E. The trend continues with en-ta, where the BLEU drops from 29.34 (ChrF++: 66.15) in cascaded to 21.34 (ChrF++: 56.07) in the E2E.

These results indicate that the cascaded approach remains advantageous for English-to-Indic translation, likely due to the mature ASR performance on English audio and the robustness of the IndicTrans2 NMT system trained on diverse high-quality parallel corpora. The modular nature of the pipeline allows each component to be fine-tuned independently, maximizing their respective capabilities.

Indic-to-English (hi-en, bn-en, ta-en): The performance gap between cascaded and E2E systems is more pronounced in the Indic-to-English direction. For hi-en, the cascaded system achieves 41.04 BLEU and 67.05 ChrF++, compared to 15.41 BLEU and 42.96 ChrF++ in the E2E track. For bn-en, the E2E model performs poorly, with only 0.459 BLEU and 14.30 ChrF++, while the cascaded model reaches 14.77 BLEU and 44.88 ChrF++. Preliminary analysis suggests that smaller amount of training data and excessive use of code-mixed language in the test set are the reason for low score for the Bengali-English pair. Similarly,

the BLEU score of ta-en model drops from 15.70 in the cascaded setup to 5.04 in the E2E setup.

The sharp decline in the E2E performance for Indic-to-English suggests that ASR on Indic audio remains a major challenge, especially in the constrained setup where access to external data or pre-trained language models is restricted. The E2E system must learn both transcription and translation jointly, which becomes challenging in low-resource settings or in speech settings consisting of code-mixed, noisy, or accented content. This highlights the difficulty of training E2E models for the Indic-origin speech, where the diversity in speech patterns and lack of rich supervised training data severely affect generalization.

5 Conclusion

Our experiments show that cascaded models still hold a strong edge in terms of accuracy and robustness, particularly in Indic to En settings while end-to-end speech translation models can be an alternative due to their simplicity and integration. With further work of using transfer learning from larger models, multilingual pre-training and data augmentation techniques such as use of synthetic data, E2E models can be at par with the cascaded models by overcoming low-resource bottlenecks in Indic languages.

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.

Karunesh Arora, Sunita Arora, and Mukund Roy. 2013. Speech to speech translation: a communication boon. *CSI Transactions on ICT*, 1.

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. *Advances in neural information processing systems*, 33:12449–12460.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Alexandre Bérard, Laurent Besacier, Ozan Caglayan, and Adrien Bardet. 2018. End-to-end automatic speech translation of audiobooks. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6224–6228. IEEE.

Liang Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5884–5888. IEEE.

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*.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Clsril-23: Cross lingual speech representations for indic languages. *Preprint*, arXiv:2107.07402.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, and Shinji Watanabe. 2020. Multilingual end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5911–5924.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, page 66.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shivam Mhaskar, Vineet Bhat, Akshay Batheja, Sourabh Deoghare, Paramveer Choudhary, and Pushpak Bhattacharyya. 2023. Vakta-setu: A speech-to-speech machine translation service in select indic languages. *arXiv preprint arXiv:2305.12518*.

Vandan Mujadia and Dipti Misra Sharma. 2023. Towards speech to speech machine translation focusing on indian languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 161–168.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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.

Anusha Prakash, Arun Kumar A, Ashish Seth, Bhagyashree Mukherjee, Ishika Gupta, Jom Kuriakose, Jordan Fernandes, K. V. Vikram, Mano Ranjith Kumar M., Metilda Sagaya Mary, Mohammad Wajahat, Mohana N, Mudit Batra, Navina K, Nihal John George, Nithya Ravi, Pruthwik Mishra, Sudhanshu Srivastava, Vasista Sai Lodagala, and 8 others. 2023. Technology pipeline for large scale cross-lingual dubbing of lecture videos into multiple indian languages. In *INTERSPEECH*, pages 3683–3684.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Elizabeth Salesky, Ramon Sanabria, Alan W. Black, and Florian Metze. 2021. Exploring low-resource speech-to-text translation across modalities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1134–1145.

Matthias Sperber and Markus 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.