Effectively combining Phi-4 and NLLB for Spoken Language Translation: SPRING Lab IITMs submission to Low Resource Multilingual Indic Track

Sankalpa Sarkar, Samriddhi Kashyap, Advait Joglekar, S. Umesh

SPRING Lab, Indian Institute of Technology Madras

Abstract

This paper presents the methodologies implemented for Spoken Language Translation for the language pairs Hindi-English, Bengali-English and Tamil-English for the Low Resource Multilingual Indic Track of The International Conference on Spoken Language Translation (IWSLT) for 2025. We adopt a cascaded approach and use a fine-tuned Phi-4 multimodal instruct model for Automatic Speech Recognition(ASR) and a fine-tuned NLLB model for Machine Translation(MT). Finally, we discuss targeted solutions (e.g. data augmentation, multilingual training, targeted fine-tuning) to boost low-resource translation, noting that significant retraining on additional Tamil data is likely needed.

1 Introduction

India is home to an incredibly diverse linguistic landscape, with over 100 officially recognized languages and thousands of dialects spoken across its vast geographic and cultural expanse. This linguistic richness reflects the countrys deep-rooted cultural heritage and regional diversity, but it also presents considerable challenges for Natural Language Processing (NLP) applications, particularly for tasks like Automatic Speech Recognition (ASR) and Machine Translation (MT).

Unlike monolingual or relatively linguistically homogeneous countries, India's multilingualism requires NLP systems to handle a wide variety of linguistic variations from phonetic and grammatical differences to script variations and region-specific vocabulary. ASR systems must account for varied accents, pronunciation patterns, and speech styles, while MT systems are required to preserve contextual accuracy and fluency across structurally and syntactically diverse language pairs.

This diversity becomes even more complex due to situations like speakers alternate between languages mid-sentence, highly inflected words and complex verb conjugations, and lexical words from one language are integrated into another. Such features are common in Indian speech and text, making the development of robust, generalizable ASR and MT models particularly challenging.

Despite these obstacles, these technologies are crucial for millions of Indians who do not speak English or other dominant languages. ASR and MT systems can help bridge the communication barrier, enabling equal access to information, digital services, and opportunities across Indias diverse population.

The "Low Resource Multilingual Indic" track proposed by IWSLT 2025 tasked participants with developing innovative methods to work with the sparse and varied resources available for three Indian languages: Hindi, Bengali, and Tamil. The participants were required to submit under various different conditions constrained or unconstrained, end-to-end or cascaded, and monolingual or multilingual. Our team participated under the unconstrained, cascaded, and monolingual category for the language pairs Hindi to English, Bengali to English, and Tamil to English. This paper outlines the implementation of our ASR and MT systems designed for these language pairs.

2 ASR

2.1 Datasets

2.1.1 SpringLab/Hindi-1482hrs

The dataset contains 1,482 hours of quality Hindi audio, and is specifically built for performing ASR tasks.

- Curated by: SPRING Lab
- Language: Hindi

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 399–404

2.1.2 AI4Bharat/SeamlessAlign [1]

BhasaAnuvaad (Jain et al., 2024), is the largest Indic-language AST dataset spanning over 44,400 hours of speech and 17M text segments for 13 of 22 scheduled Indian languages and English. This repository consists of parallel data for Speech Translation from SeamlessAlign, a subset of BhasaAnuvaad. The dataset contains 5 separate splits for different languages namely: Hindi, Tamil, Telugu, Kannada and Urdu, out of which only the Hindi split is used for training the model. Although it is an AST dataset , to perform finetuning for ASR task, we utilized the audio and the transcription column.

- Curated by: AI4Bharat
- Language: Hindi

2.1.3 SKNahin/open-large-bengali-asr-data [2]

This is a collection of publicly available ASR data for Bengali. It contains 5000 hours of audio. The dataset is divided into 9 different splits namely: commonvoice (Ardila et al., 2020), openslr (Panayotov et al., 2015), madasr, shrutilipi (Bhogale et al., 2023), kathbath (Javed et al., 2022), fleurs (Conneau et al., 2022), indictts (Conneau et al., 2022), ucla and gali, out of which only the commonvoice and ucla split were used for training. We have a filtering column called is-better to filter good-quality audio from the corpus. It is set based on the wer between original transcription and prediction taken from a Bengali-Wav2Vec2 model and word-per-second (wps).

- Curated by: SKNahin
- Language: Bengali

2.1.4 Prajwal-143/ASR-Tamil-cleaned [3]

This dataset is a combination of the Common Voice 16.0 and Open SLR datasets which is of 534 hours. It has been meticulously curated, normalized to a 16kHz sampling rate, and cleaned for better usability. This dataset aims to provide a comprehensive collection of speech data for various applications, including speech recognition, natural language processing, and machine learning research.

- Curated by: Prajwal N. Pharande
- Language: Tamil

2.2 Training

In our submission, we have fine-tuned the Phi-4 Multimodal Instruct(5.57B) (Marah Abdin, 2024) to obtain three fine-tuned models for ASR Hindi, Bengali and Tamil.

The following hyper-parameters were used during training of all the models:

- Train Epochs: 1
- Learning Rate: 4.0e-5
- Batch Size: 8
- Gradient Accumulation Steps: 4
- Optimizer: Paged Adamw 32-bit with betas=(0.9,0.999)
- LR scheduler type: cosine
- LR scheduler warmup steps: 1000
- Weight Decay: 0.01
- Max Grad Norm: 1.0

We trained the Hindi model for a total of 28263 steps on the SpringLab/Hindi-1482 hrs dataset and 31911 on the AI4Bharat/SeamlessAlign (Jain et al., 2024) dataset, the Bengali model for a total of 30113 steps on the commonvoice (Ardila et al., 2020) split and 60035 steps on the ucla split for SKNahin/open-large-bengali-asr-data dataset, finally the Tamil model for a total of 7018 steps on the Prajwal-143/ASR-Tamil-cleaned dataset. We trained the model only for 1 epoch to avoid overfitting

2.3 Evaluation

The Evaluation results for all the three models were performed on the mozilla-foundation/common-voice-17-0 (Ardila et al., 2020) [4] dataset (test split). The scores for the ASR are recorded in terms of WER and CER, listed in Table 1

[1] https://huggingface.co/datasets/ai4bharat/SeamlessAlign

[2] https://huggingface.co/datasets/SKNahin/open-large-bengali-asr-data

[3] https://huggingface.co/datasets/Prajwal-143/ASR-Tamil-cleaned

[4] https://huggingface.co/datasets/mozilla-foundation/common-voice-17-0

Language	WER	CER
Hindi	0.15156	0.06526
Bengali	0.21968	0.06288
Tamil	0.41953	0.07078

Table 1: Scores for the ASR in terms of WER and CER.

2.3.1 ASR Performance

The ASR outputs were passed to the MT stage. Errors in ASR propagate to MT, so ASR accuracy is critical. ASR errors included substitutions (e.g., similar-sounding words), deletions (inflectional suffixes), and insertions (extra syllables). Tamil had the highest WER, indicating more ASR errors.

The relatively high WER for Tamil ASR (over 0.4) can be attributed to the limited training dataonly about 200 hours were available for Tamil, in contrast to over 1,000 hours used for Hindi and Bengali. Furthermore, the Phi-4 base model used does not offer native support for Indic languages, which meant the performance for Tamil relied solely on the quality of fine-tuning.

To assess the generalization ability of our ASR models and avoid overfitting, we chose to train all language models for only one epoch. For Hindi and Bengali, which had large training sets, this approach provided sufficient exposure while maintaining regularization. Although this uniform strategy may have under exploited the Tamil dataset, it allowed us to clearly isolate data volume as a primary variable in performance.

3 MT

3.1 Datasets

3.1.1 SPRINGLab/shiksha [5]

This is a Technical Domain focused Translation Dataset for 8 Indian Languages. It consists of more than 2.5 million rows of translation pairs between all 8 languages and English.

This data has been derived from raw NPTEL documents. More information on this can be found in our paper: (Joglekar and Umesh, 2024)

[5] https://huggingface.co/datasets/SPRINGLab/shiksha

3.1.2 SPRINGLab/BPCC-cleaned [6]

A curated subset of Bharat Parallel Corpus Collection (BPCC) for 8 Indian languages. Translation pairs are filtered with LABSE score(>0.9) and further preprocessed. Useful for training high-quality translation models.

3.2 Training

For the Machine Translation of the transcriptions generated by the ASR model, we are using a fine-tuned NLLB model (3.3B) (NLLB Team, 2022) trained on the Shiksha (Joglekar and Umesh, 2024) and BPPC cleaned (English→Indic, Indic→Indic) dataset in both directions. The fine-tuned model is then used to translate transcriptions obtained through Whisper (Radford et al., 2022), Phi-4 multimodal instruct (Marah Abdin, 2024), and Data2Vec (Alexei Baevski, 2022).

Following were the training arguments that were used to fine-tune both the NLLB models:

- Learning Rate: 5e-5
- Batch Size: 8
- Weight Decay: 0.01
- Train Epochs: 5

4 Evaluation

The Evaluation results for the models were performed on the facebook/flores dataset (NLLB Team, 2022) (Goyal et al., 2021) (Guzmán et al., 2019). The scores for the MT are recorded in terms of Chrf++ and BLEU, listed in Table 2

4.0.1 Qualitative Error Analysis

Unlike purely numerical metrics like BLEU or WER, qualitative error analysis explores what kinds of errors the models make and helps identify patterns that can guide targeted improvements. We also analyzed how metric-based evaluation can sometimes fail to reflect semantic adequacy due to stylistic or lexical variation.

The figure 1 below shows a few examples of English-Hindi translation tasks where the BLEU scores are low despite reasonably acceptable translations:

These examples show that even when translations preserve meaning, surface-level metrics may penalize valid variations. Incorporating semantic

[6] https://huggingface.co/datasets/SPRINGLab/BPCC_cleaned

Source	Reference	Prediction	BLEU	chrF	Comment
Let me talk to Dad.	पहले पिताजी से बात कर लूं।	मुझे पापा से बात करने दो।	16.23	28.77	Lexical variation (formal/informal)
Yeah, I am in my second year.	हाँ, ये मेरा दूसरा साल है।	हां, मैं अपने दूसरे वर्ष में हूं।	5.52	16.96	Stylistic differences and structure shift
Am I too late?	क्या मुझे बहुत देर हो गई?	क्या मैं बहुत देर से आया हूं?	13.13	39.64	Perspective/gender difference
---	---	---	---	---	---
Come and have a seat.	এসো, এসে বোসো।	আসুন বসুন।	0	5.85	Register mismatch (formal vs. informal) --> semantically correct but stylistically different
Where are they heading to?	তারা কোথায় যাচ্ছে?	তারা কোথায় যাচ্ছে?	35.36	70.97	Metrics pick up tiny encoding differences; meaning is identical
Let us wait and watch.	চলো অপেক্ষা করি , দেখি।	আসুন অপেক্ষা করি এবং দেখি।	23.64	57.05	Both sentences are valid but differ in tone and conjunction style. informal vs. formal invitation and , vs. এবং — stylistic variation ("and")
---	---	---	---	---	---
It is a holiday.	ஸ்கூல் லீவு.	இது ஒரு விடுமுறை.	15.97	4.24	Reference has loan phrase, lexical mismatch and contextual variation
Why is it a holiday though?	ஆனால், நாளைக்கு ஏன் லீவு?	இது ஏன் ஒரு விடுமுறை?	10.4	11.42	Rephrased but semantically equivalent
Yes mom, I can't wait for it.	ஆமாம் அம்மா, நானும் அதற்காகவே காத்திருக்கிறேன்.	ஆம் அம்மா, என்னால் காத்திருக்க முடியாது.	14.54	41.95	Different idiomatic phrasing, meaning preserved, The prediction is idiomatic (lit. "I can't wait") whereas reference is literal ("I am waiting for it").

Figure 1: Sentence-level MT evaluation for Hindi, Tamil, and Bengali examples showing BLEU and chrF scores.

Language	Chrf++	BLEU
Hindi-English	0.83	0.62
Bengali-English	0.55	0.41
Tamil-English	0.79	0.85

Table 2: Datasets used for Evaluating the MT Models.

similarity metrics or human judgment could provide a more robust evaluation framework.

5 Final Evaluation

After Fine-tuning and evaluating the ASR and MT model separately, we conducted a final evaluation to test how the models were performing collectively with models namely: whisper-large-v2, data2vec-aqc[7] and NLLB. We cascaded the fine-tuned Phi-4 multimodal instruct (Marah Abdin, 2024) with NLLB and our fine-tuned NLLB model (3.3B), whisper-large-v2 (Radford et al., 2022) with NLLB (NLLB Team, 2022) and our fine-tuned NLLB model (3.3B) and data2vec-aqc with NLLB and our fine-tuned NLLB model (3.3B),

The metrics used for the evaluation were BLEU and chrf++, the results for the Bengali to English translation are listed in Table 3

6 Results

On final evaluation on the IWSLT 2025 leaderboard, with chrf as the ranking metric, we find that we perform the best on Indic to English translations, achieving first place for both Hindi to English and Bengali to English, and achieving second place for Tamil to English. However, we rank last on English to Indic evaluation. Our Evaluation

[7] https://huggingface.co/SPRINGLab/data2vec_aqc

ASR	MT	chrf++	BLEU
whisper	FT NLLB	52.5917	18.3550
whisper	NLLB	44.7125	16.7514
Data2vec-aqc	FT NLLB	53.3732	21.0471
Data2vec-aqc	NLLB	44.1628	16.3213
FT Phi-4	FT NLLB	55.2648	22.9005
FT Phi-4	NLLB	47.3048	18.3701

Table 3: Final Evaluation results for the cascaded models on Bengali to English translation task, whisper denotes the base model whisper-large-v2 (1.54B) , F denotes fine-tuned model,

results are listed in Table 4 .

6.0.1 Several factors likely contributed to the low English-Indic performance:

- Data Imbalance: While the IndicEnglish direction had sufficient high-quality training data, the EnglishIndic direction, particularly EnglishTamil, suffered from limited parallel corpora. For our MT model, we had increased the number of Indic-Indic corpora by leveraging parallel translations of the same sentence across various Indian languages. Therefore, our MT model had a much better understanding with regard to Indic languages. Same is reflected in the result, where Indic-English is performing well during evaluation. This constrained the model's ability to generalize.

- Limited Training Epochs: The ASR model was trained for only one epoch to avoid overfitting, which may have led to undertrainingespecially problematic in already low-resource settings (e.g Tamil).

Task	chrf++	BLEU
Bn-En	55.2648	22.9005
Hi-En	68.144	41.5874
Ta-En	42.0195	13.4667
En-Bn	60.8094	26.6685
En-Hi	62.3016	41.0865
En-Ta	62.3335	21.3543

Table 4: IWSLT 2025 leaderboard results for our cascaded models,

- Lack of Multilingual Transfer: Fine-tuning was done monolingually. Multilingual fine-tuning across related Indic languages could have allowed knowledge transfer and improved low-resource directions.

- Linguistic Complexity: Tamil's rich morphology and syntactic structure (e.g., SOV order) increased the difficulty of accurate translation from English without dedicated architectural or preprocessing adaptations.

Together, these limitations explain the performance gap between the two translation directions and emphasize the need for multilingual strategies and additional data in future work.

Conclusion

In this paper, we have presented our Speech Translation Systems for the low-resource Indic Languages track of IWSLT 2025 employing a cascaded ap- proach using fine-tuned models for both ASR and MT. Moving Forward we will try to employ a SLAM-ASR (Ziyang Ma, 2024) based approach to our ASR model, to get better ASR results and try to train the models more on indic languages data to generalize better.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondej 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.

Qiantong Xu Arun Babu Jiatao Gu Michael Auli Alexei Baevski, Wei-Ning Hsu. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *Machine Learning (cs.LG)*.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages. In *ICASSP*, pages 1–5. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 14 indian languages. *arXiv preprint arXiv: 2411.04699*.

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Indicsuperb: A speech processing universal performance benchmark for indian languages. *arXiv preprint*.

Advait Joglekar and Srinivasan Umesh. 2024. Shiksha: A technical domain focused translation dataset and model for indian languages. *Preprint*, arXiv:2412.09025.

Harkirat Behl Sébastien Bubeck Ronen Eldan Suriya Gunasekar Michael Harrison Russell J. Hewett Mojan Javaheripi Piero Kauffmann James R. Lee Yin Tat Lee Yuanzhi Li Weishung Liu Caio C. T. Mendes Anh Nguyen Eric Price Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Xin Wang Rachel Ward Yue Wu Dingli Yu Cyril Zhang Yi Zhang

Marah Abdin, Jyoti Aneja. 2024. Phi-4 technical report. *Computation and Language (cs.CL); Artificial Intelligence (cs.AI)*, arXiv:2412.08905. Version 1.

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 John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang (NLLB Team) NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation. *Computation and Language (cs.CL); Artificial Intelligence (cs.AI)*, arXiv:2207.04672. Version 2.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

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

Yifan Yang Zhifu Gao Jiaming Wang Zhihao Du Fan Yu Qian Chen Siqi Zheng Shiliang Zhang Xie Chen Ziyang Ma, Guanrou Yang. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.