



# End-to-End Speech Translation for Low-Resource Languages Using Weakly Labeled Data

Aishwarya Pothula<sup>1</sup>, Bhavana Akkiraju<sup>1</sup>, Srihari Bandrupalli<sup>1</sup>, Charan D<sup>1</sup>, Santosh Kesiraju<sup>2</sup>, Anil Kumar Vuppala<sup>1</sup>

<sup>1</sup>Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

<sup>2</sup>Speech@FIT, Brno University of Technology, Czechia

{aishwarya.pothula, bhavana.akkiraju}@research.iiit.ac.in

## Abstract

The scarcity of high-quality annotated data presents a significant challenge in developing effective end-to-end speech-to-text translation (ST) systems, particularly for low-resource languages. This paper explores the hypothesis that weakly labeled data can be used to build ST models for low-resource language pairs. We constructed speech-to-text translation datasets with the help of bitext mining using state-of-the-art sentence encoders. We mined the multilingual Shrutilipi corpus to build Shrutilipi-anuvaad, a dataset comprising ST data for language pairs Bengali-Hindi, Malayalam-Hindi, Odia-Hindi, and Telugu-Hindi. We created multiple versions of training data with varying degrees of quality and quantity to investigate the effect of quality versus quantity of weakly labeled data on ST model performance. Results demonstrate that ST systems can be built using weakly labeled data, with performance comparable to massive multi-modal multilingual baselines such as SONAR and SeamlessM4T.

**Index Terms:** weakly labeled data, speech translation, end-to-end models, low-resource languages

## 1. Introduction

Recent years have seen an increased interest towards building speech translation (ST) systems for low-resource languages [1, 2], enabled by transfer learning [3, 4] from large pre-trained models [5, 6], and the availability of (smaller) training datasets across several languages. The widely used datasets for ST research are CoVoST [7] and MUST-C [8], which are built on top of Mozilla common voice [9] and TED-talks respectively. The aforementioned datasets mainly span English $\leftrightarrow$ other languages, and relatively few datasets exist that are beyond English (e.g. Tamasheq $\rightarrow$ French [10], Quechua $\rightarrow$ Spanish [11]). SpeechMatrix [12], a large-scale multilingual dataset covering 136 language pairs, was created using recordings from European parliaments and expands the diversity of ST datasets by including non-English-centric translation pairs. The International Conference on Spoken Language Translation (IWSLT) is one of the long-standing, notable research communities that has been organizing an annual conference, along with shared tasks (scientific challenges) that encourage researchers and practitioners to participate in advancing state-of-the-art on speech translation technologies, spanning several domains and languages. Although some tracks release training and evaluation data with unrestricted usage of scientific research, others provide only a time-bound license (typically four months) for the data. To the best of our knowledge, there exist no freely available Indic-to-Indic speech translation datasets spanning multiple languages.

Developing speech translation systems for low-resource languages is challenging due to the lack of large and high-

quality training datasets [13]. Our research seeks to bridge this gap by building *end-to-end speech-to-text translation models* using *weakly labeled data* that are automatically curated from multilingual speech datasets. For this purpose, we use Shrutilipi (SL) [14], a corpus of multilingual speech data for 12 Indian languages. SL contains speech samples from broadcast news along with (nearly accurate) text transcripts. Using the state-of-the-art multilingual sentence encoder model (SONAR [15]), we find textual sentence pairs that are closer in the shared embedding space [16]. Pairs with high similarity score are kept for development and test sets, whereas pairs with mid-to-low similarity scores are used for training ST systems; thereby creating a weakly labeled training dataset. Sentence encoders for mining bitexts has been extensively studied and applied in the machine translation community [17–19].

Motivation for our work comes from three sources,

(i) previous work [7, 20] that relied on sentence encoders for creating CoVoST and Indic-TEDST datasets, (ii) lack of Indic-to-Indic speech translation datasets, and (iii) the research question of building ST systems from weakly labeled automatically curated data. For the last point, we employ the transfer learning approach, where our ST models are initialized from pre-trained bilingual ASR models [4]. We also use state-of-the-art multilingual massive speech-to-text models such as SONAR [15] and Seamless [5] as baseline machine and speech translation (MT, ST) systems.

To summarize:

1. We have curated an Indic to Indic speech translation dataset, dubbed *Shrutilipi-anuvaad*, that offers itself as a crucial resource for spoken translation tasks for four language pairs: Bengali-Hindi (bn-hi), Malayalam-Hindi (ml-hi), Odia-Hindi (or-hi), and Telugu-Hindi (te-hi). Details are discussed in Section 2.
2. Through our experiments (Sections 4 and 5), we demonstrate that *large amounts of weakly labeled data* can be leveraged by pre-trained ASR models, that result in ST models which perform comparably and often outperform the out-of-the-box SOTA (MT, ST) models like SONAR and Seamless.
3. Further analysis (Section 5) provides valuable insights into the *quality and quantity* of weakly labeled data and its impact on ST performance; thus helping us to understand the trade-offs involved when using available datasets. The data and code are available at: <https://github.com/aishwaryapothula/Shrutilipi-Anuvaad>.

## 2. Curating Shrutilipi-anuvaad dataset

Shrutilipi is a multilingual speech dataset derived by mining audio and text pairs from All India Radio news broadcasts [14]. We expected the presence of parallel data, since some of the

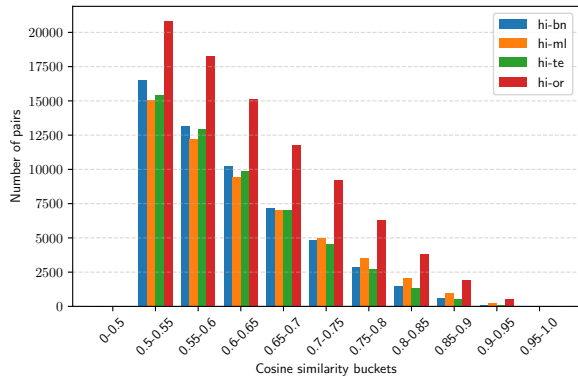


Figure 1: Histogram of cosine similarity scores computed on sentence embeddings extracted using SONAR.

news broadcasts describe the same events in multiple languages. Using SONAR text encoder, we extracted sentence embeddings for all transcripts and grouped text pairs based on cosine similarity. Pairs with a high similarity score ( $> 0.8$ ) were split equally into development and test sets; whereas pairs with mid-to-high (0.5, 0.8) similarity scores were used as weakly labeled training set. 15% of the test data has been validated by native speakers of the languages and found to be largely accurate, with the majority (more than 80%) receiving a score of 4 or 5 on a scale of 5, indicating high accuracy and alignment with human judgment. The resulting dataset contains quadruples (text pair and corresponding speech pair) for each language pair.

Table 1: Statistics of the Shrutilipi-anuvaad dataset

Lang. pair	# Utterances (hours)			
	Training $\mathcal{S}_1$	Training $\mathcal{S}_5$	Dev.	Test
hi-bn	52k (78)	8.8k (12)	1.7k (2.4)	1.7k (2.1)
hi-ml	52k (71)	7.2k (9.4)	1.0k (1.3)	1.0k (1.3)
hi-te	54k (90)	7.7k (11.6)	1.0k (1.8)	1.0k (1.6)
hi-or	81k (120)	15k (24)	3.1k (4.7)	3.1k (4.6)

Table 2: Statistics of the data for ASR training

Language	hi	bn	ml	te	or
# Utterances	662k	176k	185k	173k	225k
Hours	934	248	242	311	326

The histogram of similarity scores for four language pairs is shown in Fig. 1. We can observe that most pairs are concentrated in the similarity bin (0.5, 0.6). Pairs with score less than 0.5 are not considered to be part of ST training set, instead they were used for pre-training ASR systems. We ensured that there is no overlap of sentences between the ASR pre-training and ST dev and test sets, preventing data contamination.

Based on the similarity scores in the range (0.5, 0.8) the training set is further divided into five sets  $\mathcal{S}_1 \supset \mathcal{S}_2 \supset \mathcal{S}_3 \supset \mathcal{S}_4 \supset \mathcal{S}_5$ . The largest set  $\mathcal{S}_1$  contains all the pairs in the (0.5, 0.8) score bin, whereas the smallest set  $\mathcal{S}_5$  contains pairs from (0.7, 0.8) bin. The other splits  $\mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$  are made from bins (0.6, 0.8), (0.62, 0.8), (0.68, 0.8) respectively. The statistics of the resulting dataset is presented in Table 1. Although

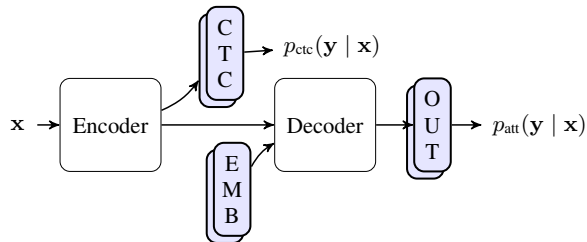


Figure 2: Encoder-decoder architecture of the bi-lingual ASR, where  $\mathbf{x}$  represents filter bank features extracted from speech signal, and  $\mathbf{y}$  represents the output text sequence. Each language has a specific CTC, embedding (EMB) and output (OUT) layers. The pre-trained bi-lingual ASR is fine-tuned for ST relying on target-language-specific CTC, EMB and OUT layers.

not given in the Table,  $\mathcal{S}_2$  is almost half the size of  $\mathcal{S}_1$  and  $\mathcal{S}_5$  is about six times smaller than  $\mathcal{S}_1$ . The average utterance lengths across all splits  $\mathcal{S}_1$ - $\mathcal{S}_5$  and the dev/test sets for all language pairs are comparable (e.g., Hi-Te:  $\mathcal{S}_1$ - $\mathcal{S}_5 = 11.4$ - $11.8$ , Dev/Test =  $11.7/10.9$ ; Hi-Bn:  $12.2$ - $13.4$ , Dev/Test =  $13.1/11.5$ ), suggesting that there is no significant length-based bias in the bucketing process. The resulting dataset is dubbed *Shrutilipi anuvaad*.

The final training data for the bilingual ASRs is created by combining sentence pairs with similarity scores between 0.5 and 0.8, along with their corresponding speech data. The left-over monolingual speech-text pairs (scores  $< 0.5$ ) are also included. Although Shrutilipi contains 12 languages, our experiments and analysis in this paper are restricted to 4 language pairs: bn-hi, ml-hi, or-hi and te-hi.

### 3. Model architectures

This section briefly explains the model architectures of ASR, ST and baseline systems used in the study. The pre-training of ASR is always bilingual and is based on transformer joint CTC/attention architecture [21], except that each language has its own CTC, embedding and output layers [4]. The architecture is depicted in Fig. 2. The ST model shares the same architecture as the ASR model, but when initialized from the pre-trained ASR, it only uses the target language-specific CTC, embedding, and output layers. This ensures that the ST model only generates text from the target language. We employed this architecture because it yielded state-of-the-art results in IWSLT’23 Marathi→Hindi low-resource ST task [1, 22].

The baseline systems are based on multilingual, multimodal foundational models, SONAR [15] and Seamless [5]. The backbone of both SONAR and Seamless is the pre-trained wav2vecbert-2.0 [23] and the NLLB [24] based massive machine translation model based on NLLB [24]. SONAR encodes both speech and text utterances into a single vector, from which the SONAR decoder generates the hypothesis in auto-regressive fashion.

### 4. Experimental setup

This section describes the details of ASR, ST and baseline models used in our experiments. The training data are augmented with speed perturbations using factors of 0.9, 1, and 1.1. From the input speech, 80-dimensional filter bank features are extracted on the fly. A sentence piece-based subword tokenizer [25] was used to build a vocabulary of 8000 tokens for each language independently. We used the ESPnet toolkit [26]

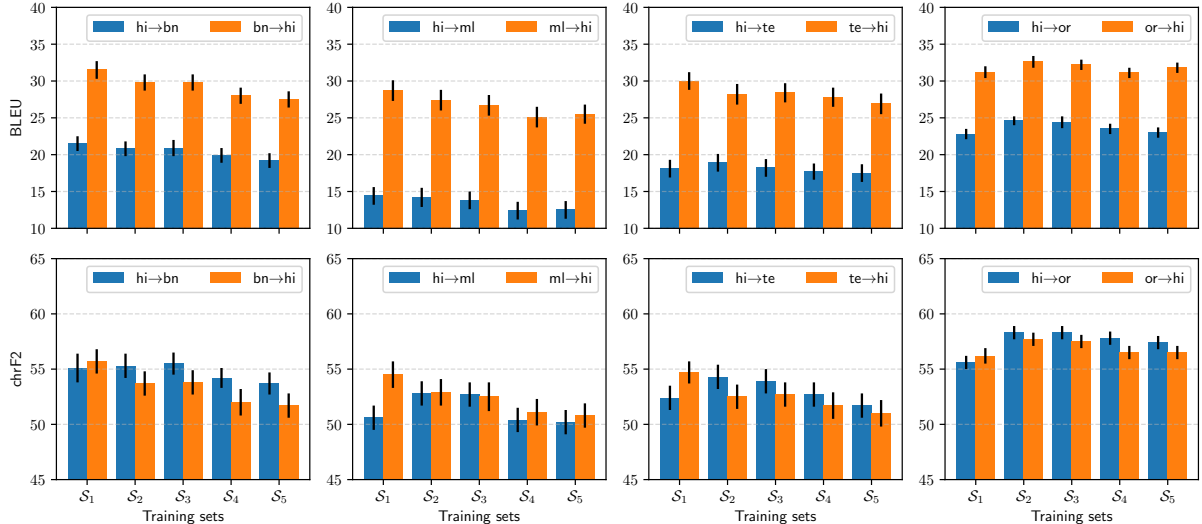


Figure 3: BLEU and ChrF2 scores along with 95% CI for 4 language pairs, varying the quality and quantity of training data ( $\mathcal{S}_1 \supset \mathcal{S}_2 \supset \mathcal{S}_3 \supset \mathcal{S}_4 \supset \mathcal{S}_5$ ). Table 4 provides the corresponding  $p$ -values for the significance tests ( $\mathcal{S}_1$  vs.  $\mathcal{S}_5$ ).

for all of our experiments, which were performed primarily on NVIDIA A100 GPUs.

#### 4.1. ASR $\rightarrow$ ST

##### 4.1.1. ASR pre-training

For ASR pre-training, we used bilingual models with joint CTC/attention objective with a CTC weight of 0.3. The training was conducted over a maximum of 100 epochs with a learning rate of 0.0005, using the Adam optimizer with gradient clipping set to 5, and a warm-up scheduler with 25,000 steps. The batch size was set to 256 with folded mini-batches. Early stopping was applied with a patience of 5 epochs, and we averaged the 10 best-performing checkpoints to get the final model. On 2 GPUs, the training took an average of 30 hours. We used joint decoding using beam search with width 10. The ASR models were evaluated using word error rates (WER). We have four bilingual ASR models, bn-hi, ml-hi, or-hi and te-hi. The word error rates are presented in Table 3. Hindi (hi) is part of each bilingual ASR system, and the reported WER is the average number. We can see from Table 3 that the Hindi-based ASR has a much lower WER compared to other ones, due to its much larger training dataset.

Table 3: WER of hi-xx bi-lingual ASR systems

bn	hi	ml	or	te
20.5	7.9	20.4	19.3	21.5

##### 4.1.2. Finetuning for ST

Speech translation models are initialized with their respective bilingual ASRs and then fine-tuned independently on each of the Shrutilipi-anuvaad training splits  $\{\mathcal{S}_1 \dots \mathcal{S}_5\}$ . The ST models are unidirectional. This transfer learning technique is used to enhance performance in low-resource scenarios. The ST models were optimized with joint CTC/attention objective for translation. The models were trained for a maximum of 100 epochs with a learning rate of  $1e-4$  and a batch size of 64. Adam opti-

mizer was used and a CTC weight of 0.3 was applied in training and decoding. Early stopping was applied with a patience of 5 epochs. Training times for ST models varied between 2 and 6 hours on a single GPU, depending on the size of the training split. We used joint decoding using beam search with width 10. ST models were evaluated using BLEU and chrF2 scores, along with 95% confidence intervals, all calculated using the sacreBLEU toolkit [27].

Table 4:  $p$  values for the statistical significance test for ST models trained on  $\mathcal{S}_1$  performing better than models trained on  $\mathcal{S}_5$  in terms of BLEU scores. This Table should be interpreted together with Fig. 3

te-hi	ml-hi	bn-hi	or-hi	hi-te	hi-ml	hi-bn	hi-or
0.002	0.002	0.002	0.011	0.048	0.001	0.002	0.159

#### 4.2. Baseline models

For our baseline, we fine-tuned Seamless (large-v2), a 2.3B parameter model. The hyperparameters used were: learning rate:  $1e-5$ , warm-up steps: 1000, max epochs: 20, patience: 7, label smoothing: 0.2, and train batch size: 5. The baseline SONAR and Seamless models without fine-tuning are also used for machine and speech translation (MT, ST) evaluation.

## 5. Results and analysis

Each of the bilingual ASRs is used to initialize a respective ST system that is trained on the splits  $\{\mathcal{S}_1 \dots \mathcal{S}_5\}$  independently. For 4 language pairs, we have 8 speech translation directions, and each ST model trained independently on 5 training splits resulted in 40 ST systems. The results of which are presented in Fig. 3. Each subplot represents ST results (BLEU scores in the first row, chrF2 in second) in both directions for a language pair. We observe that  $xx \rightarrow hi$  models (orange bars) generally achieve significantly higher BLEU and chrF2 scores compared to their reverse counterparts,  $hi \rightarrow xx$ . These trends can be at-

Table 5: Comparison of baseline MT and ST systems with our system on the test sets. **Bold** and underlined numbers indicate first and second highest BLEU scores. Parentheses show standard error representing confidence intervals.

System	Metric	bn→hi	ml→hi	or→hi	te→hi	hi→bn	hi→ml	hi→or	hi→te
SONAR MT	BLEU	24.8 (±0.9)	24.0 (±1.2)	24.6 (±0.6)	23.7 (±1.1)	11.6 (±0.7)	8.7 (±0.9)	11.7 (±0.5)	10.9 (±0.9)
	chrF2	54.2 (±1.0)	56.2 (±1.0)	55.1 (±0.5)	53.7 (±0.9)	51.3 (±1.0)	56.1 (±0.8)	51.6 (±0.5)	52.8 (±0.8)
Seamless MT	BLEU	20.0 (±0.8)	19.2 (±1.1)	22.6 (±0.6)	19.9 (±0.9)	12.1 (±0.8)	7.1 (±0.8)	11.6 (±0.5)	9.9 (±0.8)
	chrF2	51.7 (±0.7)	52.9 (±0.9)	54.7 (±0.5)	51.1 (±0.8)	52.3 (±0.7)	54.4 (±0.8)	52.1 (±0.5)	52.0 (±0.7)
SONAR ST	BLEU	15.7 (±0.7)	12.1 (±0.9)	17.6 (±0.5)	14.8 (±0.9)	7.7 (±0.5)	5.8 (±0.7)	8.2 (±0.4)	7.3 (±0.8)
	chrF2	44.2 (±0.9)	43.9 (±1.0)	48.6 (±0.5)	45.1 (±0.9)	46.0 (±0.8)	49.3 (±0.9)	46.4 (±0.5)	45.5 (±0.9)
Seamless ST	BLEU	16.6 (±0.8)	15.5 (±1.0)	18.5 (±0.5)	14.5 (±0.8)	8.1 (±0.6)	5.6 (±0.7)	10.0 (±0.5)	7.7 (±0.8)
	chrF2	40.5 (±1.2)	47.6 (±0.8)	49.1 (±0.6)	43.7 (±1.0)	43.8 (±1.0)	51.6 (±0.9)	49.0 (±0.5)	45.7 (±0.9)
Seamless ST finetuned	BLEU	<b>35.6</b> (±1.1)	<b>34.0</b> (±1.3)	<b>34.3</b> (±1.3)	<b>34.2</b> (±1.2)	<u>20.0</u> (±1.0)	<b>14.7</b> (±1.2)	<u>23.1</u> (±0.7)	<b>19.8</b> (±1.2)
	chrF2	61.0 (±0.9)	61.5 (±0.9)	60.7 (±0.5)	43.7 (±1.0)	58.3 (±1.0)	59.0 (±0.9)	59.8 (±0.5)	59.1 (±0.9)
Our ASR→ST	BLEU	<u>31.5</u> (±1.2)	<u>28.8</u> (±1.4)	<u>32.6</u> (±0.8)	30.0 (±1.2)	<b>21.5</b> (±1.0)	<u>14.4</u> (±1.2)	<b>24.6</b> (±0.6)	18.6 (±1.2)
	chrF2	55.7 (±1.1)	54.5 (±1.2)	57.7 (±0.6)	54.7 (±1.0)	55.1 (±1.3)	50.6 (±1.1)	58.3 (±0.6)	54.1 (±1.1)

tributed to the significantly larger Hindi pre-training datasets for ASR and as a consequence a stronger internal language model for Hindi. The next trend to be observed from Fig. 3 is that ST models fine-tuned on much larger but weaker sets  $\mathcal{S}_1, \mathcal{S}_2$  tend to perform better than those trained on much smaller sets of better quality  $\mathcal{S}_5$ . We conducted statistical significance tests to find the reliability of the claim. The 95% confidence intervals (CI) are depicted in the same Fig. 3, with corresponding  $p$  values tabulated in Table 4. Another observation from Fig. 3 is seen for the language pair hi-or; we see that systems trained on  $\mathcal{S}_1$  perform slightly worse than systems trained on  $\mathcal{S}_2$ , despite having twice the amount of training data, albeit poorer in quality (similarity scores between (0.5, 0.6)). This trend is not observed for other pairs, where the difference in performance between  $\mathcal{S}_1$  and  $\mathcal{S}_2$  is not statistically significant. These experiments tell us that there exists a trade-off between the quality and quantity of weakly labeled data, where beyond a point which adding more lower quality data will cause diminishing returns (e.g. hi-or).

Next, we present the results of the speech translation systems trained from scratch using the largest training split,  $\mathcal{S}_1$ . These results are shown in Table 6. It can be seen that the results are quite low, with or→hi standing out, probably due to the higher training data (Table 1). Comparison of Table 6 and Fig. 3 shows the benefits of (in-domain) pre-training.

Table 6: Results of ST systems trained from scratch on  $\mathcal{S}_1$  split.

	bn→hi	ml→hi	or→hi	te→hi
BLEU	6.8 (±0.7)	9.3 (±0.8)	14.7 (±0.5)	11.8 (±0.9)
chrF2	22.7 (±0.8)	29.4 (±1.1)	35.3 (±0.6)	31.5 (±1.0)

Table 5 compares our best ST system with SONAR and Seamless MT and ST, and Seamless fine-tuned ST systems across all language pairs. We can see from the first two rows that the SONAR MT systems perform better than the Seamless MT systems, whereas the Seamless ST system outperformed the SONAR ST systems (rows 3-4). Although these systems exhibit lower BLEU scores, their chrF2 scores perform better. This suggests that baseline models are able to correctly translate several words or subwords but struggle to translate higher-order  $n$ -grams in the right order, which explains the lower BLEU scores. The last two rows from Table 5 compares the ST system fine-tuned on the curated corpus. Seamless ST fine-tuned (row 5) achieves superior BLEU scores in 6 out of 8 directions,

while our ST model (row 6) achieves the best BLEU scores in 2 out of 8 directions. We observed that the Seamless model gave the best results when fine-tuned on the smaller but high-quality  $\mathcal{S}_5$  set, whereas our model gave best results when fine-tuned on the bigger training split  $\mathcal{S}_1$ . The performance gap is smallest for hi-bn (21.5 vs. 20.0), where our model outperforms Seamless, and largest for ml-hi (28.8 vs. 34.0), indicating that this pair benefits more from high-quality data. For Hindi-to-regional pairs (hi-bn, hi-or, hi-te, hi-ml), where translation is generally more challenging, our model remains competitive despite being trained under weaker supervision, achieving BLEU scores in a similar range compared to Seamless (14.4 vs 14.7 and 24.6 vs 23.1). These results suggest that, while fine-tuning on high-quality data leads to superior performance, large-scale weak supervision remains a viable alternative for ST in low-resource settings.

## 6. Conclusion and future work

In this paper, we aimed to train end-to-end speech translation (ST) systems by using automatically curated weakly labeled data. To this extent, we built our ST corpus, *Shrutilipi-anuvaad*, based on the multilingual Shrutilipi corpus by automatically mining parallel texts using the SONAR sentence encoder. We have conducted extensive experiments to analyze the effect of the quality and quantity of weakly labeled data on the performance of ST systems. We also observed significant improvements by employing transfer learning, i.e., initializing ST from pre-trained bilingual ASR systems. We also benchmarked and compared state-of-the-art multilingual speech translation models on the created test sets. Finally, we release the data splits for the curated corpus, there by contributing free and potentially useful datasets to the research community.

In the near future, we aim to extend the Shrutilipi-anuvaad dataset with additional languages and build ST systems for additional Indic-to-Indic directions.

## 7. Limitations

All our datasets are sourced from single-domain radio broadcasts resulting in a consistent speaking style across training and evaluation. Additionally, our pre-trained models are also derived from the same domain, which may influence performance.

## 8. Acknowledgments

Santosh Kesiraju was supported by the Ministry of Education, Youth and Sports of the Czech Republic (MoE) through the OP JAK project “Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications” (ID:CZ.02.01.01/00/23\_020/0008518). We also acknowledge Kakinada Institute of Engineering and Technology (KIET), for providing the computational resources used in this work.

## 9. References

- [1] M. Agarwal, S. Agrawal, A. Anastasopoulos, *et al.*, “Findings of the IWSLT 2023 Evaluation Campaign,” in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, ACL, 2023.
- [2] I. S. Ahmad, A. Anastasopoulos, O. Bojar, *et al.*, “Findings of the IWSLT 2024 Evaluation Campaign,” in *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, (Bangkok, Thailand (in-person and online)), pp. 1–11, ACL, Aug. 2024.
- [3] M. C. Stoian, S. Bansal, and S. Goldwater, “Analyzing ASR Pre-training for Low-Resource Speech-to-Text Translation,” in *Proc. of ICASSP*, pp. 7909–7913, IEEE, May 2020.
- [4] S. Kesiraju, M. Sarvaš, T. Pavlíček, C. Macaire, and A. Ciuba, “Strategies for Improving Low Resource Speech to Text Translation Relying on Pre-trained ASR Models,” in *Proc. of Interspeech*, pp. 2148–2152, 2023.
- [5] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elshar, H. Gong, K. Heffernan, *et al.*, “SeamlessM4T: Massively Multilingual & Multimodal Machine Translation,” 2023.
- [6] H. Inaguma, S. Popuri, I. Kulikov, P.-J. Chen, C. Wang, Y.-A. Chung, Y. Tang, A. Lee, S. Watanabe, and J. Pino, “UnitY: Two-pass direct speech-to-speech translation with discrete units,” in *Proc. of ACL*, (Toronto, Canada), pp. 15655–15680, ACL, July 2023.
- [7] C. Wang, A. Wu, J. Gu, and J. Pino, “Covost 2 and massively multilingual speech translation,” in *Interspeech 2021*, pp. 2247–2251, 2021.
- [8] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, “MuST-C: a Multilingual Speech Translation Corpus,” in *Proc. of NAACL, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 2012–2017, ACL, June 2019.
- [9] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, (Marseille, France), pp. 4218–4222, European Language Resources Association, May 2020.
- [10] M. Zanon Boito, F. Bougares, F. Barbier, S. Gahbiche, L. Barrault, M. Rouvier, and Y. Estève, “Speech resources in the Tamasheq language,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (Marseille, France), pp. 2066–2071, European Language Resources Association, June 2022.
- [11] R. Cardenas, R. Zevallos, R. Baquerizo, and L. Camacho, “Siminchik: A speech corpus for preservation of southern Quechua,” *ISI-NLP 2*, p. 21, 2018.
- [12] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswami, C. Wang, J. Pino, B. Sagot, and H. Schwenk, “SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations,” in *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, (Toronto, Canada), pp. 16251–16269, ACL, July 2023.
- [13] Y. Jia, Y. Ding, A. Bapna, C. Cherry, Y. Zhang, A. Conneau, and N. Morioka, “Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation,” in *Interspeech 2022*, pp. 1721–1725, 2022.
- [14] K. Bhogale, A. Raman, T. Javed, *et al.*, “Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [15] P.-A. Duquenne, H. Schwenk, and B. Sagot, “SONAR: Sentence-Level Multimodal and Language-Agnostic Representations,” 2023.
- [16] P.-A. Duquenne, H. Gong, and H. Schwenk, “Multimodal and multilingual embeddings for large-scale speech mining,” in *Neural Information Processing Systems*, 2021.
- [17] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, “CCMatrix: Mining billions of high-quality parallel sentences on the web,” in *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 6490–6500, ACL, Aug. 2021.
- [18] S. Sloto, B. Thompson, H. Khayrallah, T. Domhan, T. Gowda, and P. Koehn, “Findings of the WMT 2023 shared task on parallel data curation,” in *Proceedings of the Eighth Conference on Machine Translation*, (Singapore), pp. 95–102, ACL, Dec. 2023.
- [19] V. Chaudhary, Y. Tang, F. Guzmán, H. Schwenk, and P. Koehn, “Low-resource corpus filtering using multilingual sentence embeddings,” in *Proceedings of the Fourth Conference on Machine Translation (Volume 3)*, (Florence, Italy), pp. 261–266, ACL, Aug. 2019.
- [20] N. Sethiya, S. Nair, and C. Maurya, “Indic-TEDST: Datasets and baselines for low-resource speech to text translation,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, (Torino, Italia), pp. 9019–9024, ELRA and ICCL, May 2024.
- [21] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [22] S. Kesiraju, K. Beneš, M. Tikhonov, and J. Černocký, “BUT systems for IWSLT 2023 Marathi - Hindi low resource speech translation task,” in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, (Toronto, Canada (in-person and online)), pp. 227–234, ACL, July 2023.
- [23] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, “w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250, 2021.
- [24] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heffernan, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, *et al.*, “No Language Left Behind: Scaling Human-Centered Machine Translation,” *arXiv*, vol. abs/2207.04672, 2022.
- [25] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” in *Proc. of the 2018 Conference on EMNLP: System Demonstrations*, (Brussels, Belgium), pp. 66–71, ACL, Nov. 2018.
- [26] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, “ESPnet-ST: All-in-one speech translation toolkit,” in *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*, pp. 302–311, ACL, July 2020.
- [27] M. Post, “A Call for Clarity in Reporting BLEU Scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Brussels, Belgium), pp. 186–191, ACL, Oct. 2018.