Indic-TEDST: Datasets and Baselines for Low-Resource Speech to Text Translation

Nivedita Sethiya[1], **Saanvi Nair**[2]* **, Chandresh Kumar Maurya**[1]
[1]Indian Institute of Technology Indore, Simrol, Indore, Madhya Pradesh, India 453552
[2]MS Ramaiah Institute of Technology, Bengaluru, Karnataka, India 560054
{phd2201201003, chandresh}@iiti.ac.in, 1ms21cs110@msrit.edu

Abstract

Speech-to-text (ST) task is the translation of speech in a language to text in a different language. It has use cases in subtitling, dubbing, etc. Traditionally, ST tasks have been solved by cascading automatic speech recognition (ASR) and machine translation (MT) models which leads to error propagation, high latency, and training time. To minimize such issues, end-to-end models have been proposed recently. However, we find that only a few works have reported results of ST models on a limited number of low-resource languages. To take a step further in this direction, we release datasets and baselines for low-resource ST tasks. Concretely, our dataset has 9 language pairs and benchmarking has been done against SOTA ST models. The low performance of SOTA ST models on Indic-TEDST data indicates the necessity of the development of ST models specifically designed for low-resource languages.

Keywords: Speech-to-text Translation, Low-Resource Languages, Video Subtitling, Automatic Speech Recognition, Machine Translation, Indic Languages

1. Introduction

In the realm of spoken language processing, Speech-to-text Translation (ST) stands as a critical subtask that resides at the intersection of natural language processing. The primary objective of ST is to convert speech in one language into written text in another language. Traditionally, this task has been carried out by human language translators possessing proficiency in both the source and target languages. However, the availability of translators with expertise in multiple languages is limited. Consequently, there is a pressing necessity for a specialized model designed to excel in the specific domain of ST tasks across different languages.

To facilitate the development of such a model, it is essential to have parallel data comprising speech in the source language, corresponding transcription (optional), and translation text in the target language. While the datasets for ST are readily available for high-resource languages, they remain scarce for low-resource languages. Recent efforts have witnessed substantial progress in the domain of ST, primarily focused on high-resource languages, leaving low-resource languages with a considerable journey ahead. This imbalance is largely attributed to the limited availability of data for low-resource languages, as most deep-learning models heavily rely on data abundance.

The collection of such data for low-resource languages proves to be a challenging endeavor. Despite the extensive work conducted on ST for high-resource languages, overall efficiency still falls short of expectations (Le et al., 2023; Wang et al., 2022; Ye et al., 2021). Several contributing factors include the insufficient volume of training data, variations in speaker accents, age demographics of the speakers, noise interference in the speech data, and numerous other complex factors. Furthermore, the current ST models have demonstrated sub-optimal performance when applied to low-resource languages, highlighting substantial room for enhancement in this domain (Li et al., 2020a; Ouyang et al., 2022).

While extensive research has been conducted on ST translation across various language families, there remains a notable gap in the exploration of this domain for low-resource Indian languages. Surprisingly, no datasets exist specifically tailored for the ST task in Indian languages, encompassing neither the Dravidian nor the Devnagri language families. In this study, we present a dataset for ST in nine low-resource Indian languages from the TED website[1], named Indic-TEDST[2]. In the Indic-TEDST dataset, speech and transcriptions are in English, while the translation texts are in 9 Indian languages. The dataset curation incorporates TED data collected prior to August 2023. The Indian languages covered in the dataset are Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Punjabi (pa), Tamil (ta), and Telugu (te).

* Work done while interning at Indian Institute of Technology Indore

[1]https://www.ted.com/

[2]Will be available at https://github.com/Nivedita5/Indic-TEDST

LREC-COLING 2024, pages 9019–9024

The paper is structured as follows: In Section 2, we examine existing datasets related to the ST task. Section 3 delves into the Indic-TEDST dataset, encompassing discussions about its resources (§3.1), data collection and preprocessing (§3.2), cross-lingual alignment techniques (§3.3), speech-text alignment methods (§3.4), dataset structure (§3.5), and dataset statistics (§3.6). Moving to Section 4, we explore state-of-the-art models on which the dataset is trained and introduce baseline models. In Section 5, we present the results achieved by these models on the Indic-TEDST dataset. Finally, Section 6 summarizes and concludes the research presented in the paper.

2. Existing ST Data

In a typical ST corpus, it is customary to find spoken language in one dialect paired with written text in another. Consequently, various authors gathered synthetic data for this purpose from ASR datasets like WSJ[3], BTEC[4], Miami (Deuchar, 2008), CHiME-4 (Christensen et al., 2010), and MSLT (Federmann and Lewis, 2016), and then meticulously refined them to fit the context of ST using MT techniques. Despite the existence of several synthetic datasets, considerable efforts have been dedicated to the creation of more human-like, organic ST corpora. These include augmented Librispeech (Kocabiyikoglu et al., 2018), MuST-C (Di Gangi et al., 2019), Europarl-ST (Iranzo-Sánchez et al., 2020), and Voxpopuli (Wang et al., 2021) have introduced extensive resources for high-resource languages. However, these datasets have primarily concentrated on European languages, leaving a discernible void in the realm of ST for other linguistic families. Recognizing this gap, the Kosp2e (Cho et al., 2021) dataset emerged as a solution, presenting a comprehensive Korean ST corpus. Similarly, GigaST (Ye et al., 2022) unveiled an ST corpus that encompasses Chinese and German translations, broadening the scope of ST datasets beyond European languages.

The existing datasets exhibit a notable deficiency in the inclusion of a wide array of language families. Consequently, the field of ST has experienced limited exploration when it comes to Indian languages, characterized as low-resource languages.

[3] https://catalog.ldc.upenn.edu/LDC93s6a

[4] http://universal.elra.info/product_info.php?cPath=37_39&products_id=80

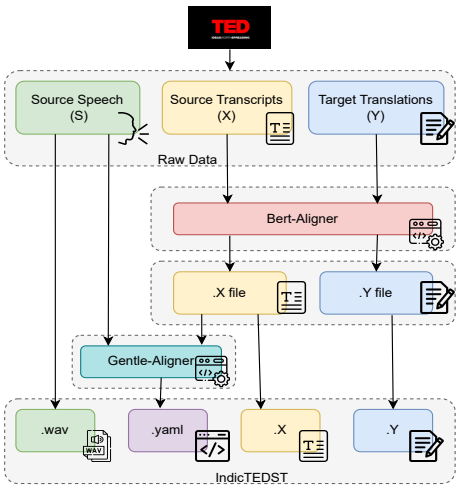

Figure 1: process diagram followed to curate the Indic-TEDST dataset. S and X denote the speech and text in the source language, respectively, and Y denotes the text in the target language.

3. Materials and Methods

3.1. Resource

The Indic-TEDST dataset is derived from the TED website, a platform dedicated to hosting talks in the English language, where individuals from diverse backgrounds share their insights and experiences spanning various domains. These talks typically vary in duration, ranging from 3 to 18 minutes, and are accompanied by transcriptions and translations in multiple languages. There exists an ST dataset known as MuST-C, sourced from TED talks.

3.2. Data Acquisition and Preprocessing

The process of data acquisition involves web scraping of the TED website, where we sourced the raw data, as can be seen from Figure 1. In this context, the English TED talk videos served as the source language speech. Additionally, we collect the English language transcriptions and the Indian language translations of these talks from the website. The audio data, representing the speech, is converted into the WAV format, a common digital audio file format. Concurrently, we execute a preprocessing step at the sentence level for cross-lingual text alignment.

The preprocessing involves preparing transcripts from TED Talks, which includes addressing unicode/encoding issues and splitting sentences based on strong punctuation marks. For English sentences, the data is split and organized using punctuation marks, and non-English content is removed. Indian language sentences undergo a customized script for line-wise splitting based on language-specific punctuation, addressing encod-

ing issues and ensuring alignment with English text using BERTALIGN. Audio files are converted to $16kHz$ with the required bit rate, transformed from mp4 to wav, and filtered for speech-only content in English, excluding music and non-English speech files.
We explored Montreal forced aligner, Prosodylab-Aligner, etc., for cross-lingual alignment and Gargantua, BlueAlign, etc to align speech and text in English. The performance of all these models was found unsatisfactory on manual validation. Finally, we found Bert Align for cross-lingual alignment of texts and Gentle Aligner for speech-text alignment perform satisfactorily.

3.3. Cross-Lingual Alignment

The cross-lingual sentence alignment is achieved with the help of the Bert Align tool[5], a sophisticated linguistic alignment tool. Bert Alginer compares sentences in the source language (English) and the target Indian language transcripts, aligning them in a coherent manner. The result is the creation of a single text file where the aligned lines alternate between English and the respective Indian languages. This alignment procedure serves as a fundamental building block for subsequent linguistic analysis.

3.4. Speech-Text Alignment

Another critical subtask of speech-text alignment is synchronizing the speech and text components. Leveraging the Gentle Forced Aligner (Ochshorn and Hawkins, 2017), a dedicated aligner tool, we generate the .WAV files with their corresponding transcripts. The alignment allows us to create a dataset available in both CSV and JSON formats. These files serve as a crucial resource in unraveling the intricate relationship between spoken words and their corresponding textual representations.

3.5. Data Organization and Directory Structure of Indic-TEDST

The dataset is organized within a directory structure known as Indic-TEDST. For each specific Indian language, a dedicated folder named "Indic-TEDST-lang" is created, with *lang* representing the particular Indian language. As illustrated in Figure 2, the directory structure for the Hindi language, denoted as Indic-TEDST-Hindi within Indic-TEDST, is exemplified. Within this folder, a subfolder named "en-lang" is created, housing another subfolder named *data*. These *data* subfolders have categories: *test*, *train*, and *valid*, which further feature two subfolders, *text* and *wav*. The *text* subfolder includes three distinct files: one encompassing the English transcription (labeled as .en), another containing the transcription in the Indian language (identified as .lang), and finally, a .yaml file enriched with essential parameters, facilitating the mapping of .wav files to their corresponding .en files. Simultaneously, the *wav* subfolder contains all the associated .wav files, which are cross-referenced in the .yaml file for comprehensive dataset organization.

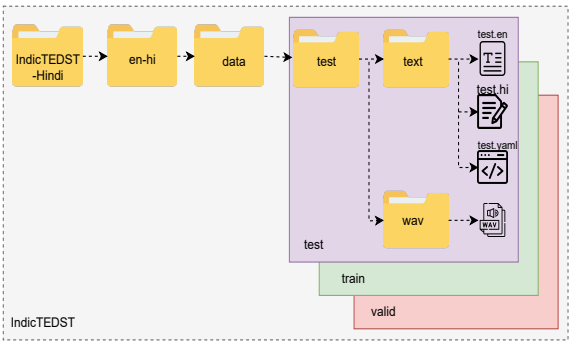

Figure 2: Directory Structure of the Indic-TEDST dataset. Here, the dataset is shown for the Hindi language. The hierarchy of the directory will be the same for test, train, and valid, inside the data folder.

3.6. Statistics of Indic-TEDST

The Indic-TEDST dataset encompasses an ST corpus featuring 9 low-resource Indian languages. As depicted in Table 1, uniformity is maintained across all the corpora, with an equivalent number of lines in their .en, .lang, and .yaml files. However, due to the inherent linguistic disparities, the number of tokens in the .en and .lang files is not uniform. The count of audio files corresponds to the number of distinct talks, each delivered by an individual speaker. Additionally, the speech hours denote the cumulative duration of speech in a given language. Each of these parameters is meticulously categorized into test, train, and valid subsets, thereby establishing a comprehensive and structured dataset.

4. Experiments

We conducted training for the end-to-end Speech Translation tasks on all languages included in the Indic-TEDST dataset. Our experimentation encompassed state-of-the-art models, specifically the Transformer (Chan et al., 2016; Vaswani et al., 2017) and Conformer (Gulati et al., 2020), utilizing all available Indic-TEDST datasets. Additionally, we explored training these models with the incorporation of an ASR encoder that had been pre-trained on English MuST-C ASR data of 430 hours. These experiments were carried out using the FairSeq toolkit (Ott et al., 2019), and the

[5] https://github.com/bfsujason/bertalign

Lang	#Lines			#Tokens(en)			#Tokens(lang)			#Audio files			#Speech (Hrs)		
en→	test	train	valid	test	train	valid	test	train	valid	test	train	valid	test	train	valid
bn	1.1	5.1	1.3	19.3	89.4	22.1	17.3	80.4	20.4	15	106	30	2.09	9.20	2.30
gu	0.5	4.5	1.3	10.1	76.9	22.6	10.2	76.8	22.7	14	91	26	1.09	7.95	2.23
hi	7.2	45.8	7.6	118.6	752.6	130.3	138.0	890.5	158.5	75	528	150	13.52	76.46	13.52
kn	0.2	0.9	0.2	2.9	13.8	2.9	2.0	10.2	2.1	2	18	4	0.33	1.46	0.27
ml	0.4	2.2	1.1	6.7	37.1	19.1	4.7	25.8	12.8	7	48	13	0.69	3.74	1.77
mr	0.7	10.5	2.7	12.2	179.2	40.2	10.6	153.8	35.1	22	153	49	1.31	18.45	4.09
pa	0.2	0.5	0.1	2.0	8.1	1.0	2.3	9.4	1.2	1	6	1	0.23	0.85	0.08
ta	2.2	8.0	2.1	38.9	135.1	35.4	28.0	101.5	27.3	20	145	42	4.04	14.41	3.56
te	0.3	2.1	0.7	5.4	36.1	10.8	4.3	27.9	8.3	7	50	14	0.56	3.69	1.13

Table 1: Statistics of Indic-TEDST dataset. #Lines and #Tokens (.en & .lang) are in terms of thousands(K). All the data in the above table is approximated.

Lang	Transformer		Conformer	
en→	w/o ASR	with ASR	w/o ASR	with ASR
bn	1.84	5.86	1.30	5.07
gu	2.75	3.50	2.60	3.60
hi	2.48	2.72	5.23	2.60
kn	0.50	2.90	0	2.53
ml	0	2.92	1.80	3.50
mr	4.20	4.65	2.24	4.63
pa	0	1.56	0	1.62
ta	0.40	0.50	1.90	1.20
te	2.70	1.20	1.85	0.90

Table 2: BLEU score of various models on Indic-TEDST dataset

model training was executed on a single NVIDIA GeForce RTX A5000 GPU equipped with 16GB of VRAM. The following combinations of training were applied to the Indic-TEDST dataset:

- Transformer without ASR encoder
- Transformer with pre-trained ASR encoder
- Conformer without ASR encoder
- Conformer with pre-trained ASR encoder

Both the Transformer and Conformer models in this study utilize default settings from the Fairseq toolkit. The Transformer configuration includes an input embedding dimension of 256, 12 encoder layers, 6 decoder layers, a hidden dimension of 2048 for feedforward sub-layers, 4 attention heads, and a ReLu activation function. Meanwhile, the Conformer configuration comprises a similar input embedding dimension, 16 encoder layers, 1 decoder layer, a hidden dimension of 2048 for feedforward sub-layers, 4 attention heads for the encoder, 8 attention heads for the decoder, and an activation function of ReLu.
Table 2 presents the results obtained across the test subset of the specific languages within the Indic-TEDST dataset. These results are denoted by BLEU scores (Papineni et al., 2002), which were calculated using the scarebleu library within the FairSeq toolkit. We provide the average of all the BLEU scores obtained with a beam size of 5 for each language.

5. Results

As detailed in Table 2, it's evident that the BLEU scores for certain languages are notably low when neither the Transformer nor the Conformer models are equipped with a pre-trained ASR encoder. This can be primarily attributed to the limited size of data available for those specific languages, categorizing them as very low-resource languages. In the broader context, when comparing the performance of both models trained with and without a pre-trained ASR encoder, a significant disparity in BLEU scores becomes apparent. Despite the provided BLEU scores in Table 2, it is apparent that there is still substantial room for improvement in the context of the Indian low-resource languages encompassed by the Indic-TEDST dataset. The BLEU scores on the Indic-TEDST dataset are relatively low and fall below the theoretical standard for acceptability. This is found consistent with the trend seen in various low-resource languages having BLEU score less than 10, such as Fa→En, Mn→En, Id→En, etc (Li et al., 2020b).

The low results across certain languages can be attributed to three primary factors. First, there is a scarcity of data, particularly evident in languages like Kannada, Punjabi, and Telugu. Second, the quality of data, specifically translations, is another significant factor. For example, in the case of Hindi and Tamil, there is often only one Tamil sentence available for every 10-15 English sentences. Certainly, once these factors are addressed, the model is expected to enhance its performance. Lastly, upon examining the results, it becomes apparent that the inferior performance of the ST models stems from the decoder being trained solely on the Indic-TEDST data, while the encoder is pre-trained on ASR data. We believe that one can get better results if pre-training of the decoder is done on a large number of monolingual corpus by optimizing the MT loss, which will be a future work where BLUE scores of the models tested will serve as the baseline. Also, by collecting more data, data augmentation might help.

6. Conclusion

This paper introduces Indic-TEDST, an ST dataset specifically designed for low-resource Indian languages. The dataset comprises ST data in 9 Indian languages, all of which are paired with English speech, transcriptions, and parallel translation texts in the respective Indian languages. State-of-the-art models have been trained on these datasets, serving as the baseline results for this corpus. However, the initial results indicate that there is substantial room for improvement, likely due to the limited quantity of data currently available. In the future, our plans involve expanding the dataset by incorporating more diverse data sources and content across these languages. Furthermore, we intend to introduce additional low-resource Indian languages to further enrich the ST task.

7. Bibliographical References

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. Kosp2e: Korean speech to english translation corpus. *arXiv preprint arXiv:2107.02875*.

Heidi Christensen, Jon Barker, Ning Ma, and Phil D Green. 2010. The chime corpus: a resource and a challenge for computational hearing in multisource environments. In *Eleventh annual conference of the international speech communication association*.

Margaret Deuchar. 2008. The miami corpus: Documentation file. *Bangortalk, bangortalk. org. uk/docs/Miami_doc. pdf*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

Christian Federmann and William Lewis. 2016. Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *Proceedings of the 13th International Conference on Spoken Language Translation*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. *arXiv preprint arXiv:1802.03142*.

Hang Le, Hongyu Gong, Changhan Wang, Juan Miguel Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: Ctc meets optimal transport. *ArXiv*, abs/2301.11716.

Xian Li, Changhan Wang, Yun Tang, C. Tran, Yuqing Tang, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020a. Multilingual speech translation from efficient finetuning of pretrained models. In *Annual Meeting of the Association for Computational Linguistics*.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020b. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

RM Ochshorn and Max Hawkins. 2017. Gentle forced aligner. *github. com/lowerquality/gentle*.

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. *arXiv preprint arXiv:1904.01038*.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. Waco: Word-aligned contrastive learning for speech translation. *arXiv preprint arXiv:2212.09359*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Pro-*

ceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2022. Simple and effective unsupervised speech translation. *arXiv preprint arXiv:2210.10191*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2022. Gigast: A 10,000-hour pseudo speech translation corpus. *arXiv preprint arXiv:2204.03939*.