SRI-B's systems for IWSLT 2023 Dialectal and Low-resource track: Marathi-Hindi Speech Translation

Balaji Radhakrishnan, Saurabh Agrawal, Raj Prakash Gohil, Kiran Praveen,
Advait Vinay Dhopeshwarkar, Abhishek Pandey
Samsung R&D Institute,Bangalore
{balaji.r, saurabh.a, raj.gohil, k.praveen.t, a.dhopeshwar, abhi3.pandey}@samsung.com

Abstract

This paper describes the speech translation systems SRI-B developed for the IWSLT 2023 Evaluation Campaign Dialectal and Low-resource track: Marathi-Hindi Speech Translation. We propose systems for both the constrained (systems are trained only on the datasets provided by the organizers) and the unconstrained conditions (systems can be trained with any resource). For both the conditions, we build end-to-end speech translation networks comprising of a conformer encoder and a transformer decoder. Under both the conditions, we leverage Marathi Automatic Speech Recognition (ASR) data to pre-train the encoder and subsequently train the entire model on the speech translation data. Our results demonstrate that pre-training the encoder with ASR data is a key step in significantly improving the speech translation performance. We also show that conformer encoders are inherently superior to its transformer counterparts for speech translation tasks. Our primary submissions achieved a BLEU% score of 31.2 on the constrained condition and 32.4 on the unconstrained condition. We secured the top position in the constrained condition and second position in the unconstrained condition.

1 Introduction

Speech translation (ST) is the task of automatically translating a speech signal in a given language into text in another language. While rapid strides have been made in speech translation in recent times, this progress has been restricted to a small number of high resource languages. This progress excludes sizable sections of people who speak languages that have very little speech data available. So, for these speech systems to be beneficial and impactful in the real world, they have to be developed and shown to work on low-resource languages as well.

In order to mitigate these issues and encourage research on low-resource languages, IWSLT propose a dialectal and low-resource speech translation track (Agarwal et al., 2023) as a part of their 2023 shared tasks evaluation campaign. While this track includes various low resource languages, we focus our efforts on the Marathi-Hindi language pair. The goal of this task is to translate Marathi speech to it's corresponding Hindi text. Marathi and Hindi are both Indo-Aryan languages used in India. Even though there were 83 million people across India speaking Marathi as per the 2011 census of India, it lacks sufficient speech data to support modern speech translation systems.

This paper discusses our work and submissions on the Marathi-Hindi low-resource speech translation task. Our experiments in this paper focus only on end-to-end architectures. We begin our experiments with a simple end-to-end Transformer and build on this approach with the following key contributions that significantly better our final performance:

- Encoder pre-training with Marathi ASR data.
- Replacing the Transformer encoder blocks with Conformer encoder blocks.
- Utilizing the dev split during speech translation training for the final submissions.

2 Related work

Traditionally, speech translation was performed using cascaded systems (Ney, 1999) (Casacuberta et al., 2008) (Post et al., 2013) (Kumar et al., 2014) of ASR and Machine Translation (MT) models. In this approach, speech was first transcribed using an ASR model and then the transcriptions were translated to text in the target language with the help of a MT model. This approach however possessed several key drawbacks like error propagation, increased latency, and architectural complexity due to multiple models.

The first attempt towards building an end-to-end speech translation system was by (Bérard et al., 2016), where they built a system that eliminated

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 449–454
July 13-14, 2023 ©2023 Association for Computational Linguistics

Type	#Utterances	Hours
train	7990	15.53
dev	2103	3.39
test	2164	4.26

Table 1: Details of speech translation (ST) data.

the need for source language transcriptions. Similarly, (Weiss et al., 2017) proposed an attention based encoder-decoder architecture for end-to-end speech translation that exhibited improved performance over cascaded systems. (Bentivogli et al., 2021) perform a detailed comparison between the paradigms of cascaded and end-to-end speech translation.

Developing speech translation systems for low-resource scenarios are especially challenging given the scarcity of training data. Speech translation systems submitted in IWSLT 2019 (Niehues et al., 2019) tended to prefer cascaded approaches for low-resource tracks. The cascaded approach which was favoured in (Le et al., 2021), used a hybrid ASR system with wav2vec features followed by a MT model for two low-resource language pairs. Recently, as system trained with joint optimization of ASR, MT and ST (Anastasopoulos et al., 2022) exhibited good performance. Also, usage of self-supervised learning based pre-trained models such as XLR-S (Babu et al., 2021) and mBART (Tang et al., 2020) have been shown to be effective, especially for low-resource scenarios.

3 Data description

The challenge data consists of Marathi speech to Hindi text translation data from the news domain for the model training and development which we shall henceforth refer to as ST (speech translation) data. The details of this dataset has been mentioned in Table 1. This dataset was directly shared with all the participants involved. Development *(dev)* and test *(test)* sets were also provided for assessing the model performance. Hindi text labels for the test set were kept blind for all the participants.

The organizers shared additional Marathi audio data along with its transcripts which can be used for the constrained condition, the details of which have been mentioned in Table 2. Common Voice (Ardila et al., 2019) is a publicly available multi-language dataset prepared using crowd-sourcing. OpenSLR (He et al., 2020) is a multi-speaker speech corpora intended for text-to-speech (TTS) applications. Indian Language Corpora (Abraham et al., 2020) consists of crowd-sourcing recordings of low-income workers. From all three datasets, only the Marathi language subsets were utilized for training purposes.

For the unconstrained condition, in addition to the aforementioned datasets, IIIT-H Voices (Prahallad et al., 2012) (Prahallad et al., 2013) and IITM Indic TTS (Baby et al., 2016) were also utilized, both of which were designed for building TTS systems.

4 System Description

All the models we trained for this challenge are end-to-end speech translation (ST) systems. For the purposes of this challenge, we tried two architectures: Listen, attend and spell (LAS) (Chan et al., 2016) style Transformer (Vaswani et al., 2017) and the same model with its encoder replaced with Conformer (Gulati et al., 2020) layers. Both the models were implemented using the Fairseq S2T toolkit (Ott et al., 2019).

The Conformer model consists of a 16-layer Conformer encoder paired with a 6-layer Transformer decoder. The Transformer model comprises of a 12-layer Transformer encoder and a 6-layer Transformer decoder. In all the cases where pre-training is involved, the encoder blocks are pre-trained (Bahar et al., 2019) using Marathi ASR data mentioned in *Table 2*. Then, the model is trained on the Marathi-Hindi ST data with the encoder initialized from the previous ASR pre-training stage. Relative positional encoding was used in the case of the Conformer model.

For speech inputs, 80-channel log mel-filter bank features (25ms window size and 10ms shift) were extracted with utterance-level CMVN (Cepstral Mean and Variance Normalization) applied. SpecAugment (Park et al., 2019) is applied on top of this feature set. We experimented with character vocabulary and a 1000 BPE (Byte Pair Encoding) vocabulary and found that the former performs better for our task.

Adam (Kingma and Ba, 2014) with a learning rate of 2×10^{-3} was the optimizer of choice for all the experiments. Inverse square-root scheduling available in the toolkit was used with a warm-up of 1000 steps. Label-smoothed-cross-entropy with 0.1 as label smoothing was used as the criterion across all the experiments. We set dropout (Srivastava et al., 2014) to 0.15 during ASR pre-training and

Dataset	Condition	Hours
Indian Language Corpora	Constrained	109
Common Voice	Constrained	3.7
OpenSLR	Constrained	3
IIIT-H Voices	Unconstrained	40
IITM Indic TTS	Unconstrained	20

Table 2: Details of Marathi ASR datasets used for pre-training.

0.1 during ST training. We pre-train on the ASR data for 6000 steps and then train on the ST data for 2250 steps. After ST training, we average the last 10 checkpoints to create the final model. We used a beam size of 10 for decoding.

4.1 Constrained condition

For the constrained condition, we are only permitted to use the data provided by the organizers. For the constrained models, wherever pre-training is involved, we only utilize the 3 constrained datasets from *Table 2*. For this condition, we train the following models:

- The Transformer model trained with only the train split from the ST data.
- The Conformer model trained with only the train split from the ST data.
- The Transformer model encoder pre-trained with constrained ASR data mentioned in *Table 2* and then trained with only the train split from the ST data.
- The Conformer model encoder pre-trained with constrained ASR data mentioned in *Table 2* and then trained with only the train split from the ST data. This served as our constrained contrastive model for the final submission.
- The Conformer model encoder pre-trained with constrained ASR data mentioned in *Table 2* and then trained with both the train and the dev splits from the ST data. This served as our constrained primary model for the final submission.

4.2 Unconstrained condition

For the unconstrained condition, wherever pre-training is involved, we utlize all of the datasets mentioned in *Table 2*, both constrained and unconstrained. Since the Conformer models outperform the Transformer ones as can be gleaned from *Table 3*, we chose to use only Conformer models for the unconstrained condition. We train the following models for the unconstrained condition:

- The Conformer model encoder pre-trained with constrained and unconstrained ASR data mentioned in *Table 2* and then trained with only the train split from the ST data.This served as our unconstrained contrastive model for the final submission.
- The Conformer model encoder pre-trained with constrained and unconstrained ASR data mentioned in *Table 2* and then trained with both the train and the dev splits from the ST data. This served as our unconstrained primary model for the final submission.

5 Results

The results for all the models we trained can be seen in *Table 3*. The first striking result is that, irrespective of the scenario, the Conformer encoder strongly outperforms the Transformer encoder. Replacing the Transformer encoder blocks with it's Conformer counterpart results in the dev split BLEU score increasing by 3.2 points. Conformers are already state of the art when it comes to speech recognition, so it would make inherent sense that this advantage would carry over to speech translation as well.

Encoder pre-training with Marathi ASR data also results in a significant improvement in speech translation performance.This is a commonly used strategy while training speech translation models and allows us to increase the BLEU score on the dev split by 8.5 and 11.8 points on the Transformer and Conformer models respectively. Two additional Marathi ASR datasets were added for pre-training the encoder in the unconstrained condition. This resulted in the BLEU score increasing by 4.1 points on both the dev and test splits.

Condition	Model	Pretraining	Training Data		Dev	Test	
			ASR	ST	BLEU(%)	BLEU(%)	CHRF2(%)
Constrained	Transformer	✗	–	train	1.02	–	–
Constrained	Conformer	✗	–	train	4.26	–	–
Constrained	Transformer	✓	constrained	train	9.55	–	–
Constrained	Conformer	✓	constrained	train	16.09	25.7	49.4
Constrained	Conformer	✓	constrained	train+dev	–	31.2	54.8
Unconstrained	Conformer	✓	all	train	20.22	29.8	53.2
Unconstrained	Conformer	✓	all	train+dev	–	32.4	55.5

Table 3: Results for all our trained models on dev & test splits. Here *all* indicates that both constrained and unconstrained datasets were used for ASR pretraining.

Finally, since the dev and test splits come from a similar distribution, including the dev split in speech translation training boosted our BLEU scores on the test split by 5.5 and 2.6 points in the cases of constrained and unconstrained conditions respectively. Utilizing the dev split for speech translation training also narrowed down the gap in performance between the unconstrained and constrained models on the test split.

6 Conclusion

In this paper we present our approaches to the IWSLT 2023 Evaluation Campaign Dialectal and Low-resource track: Marathi-Hindi Speech Translation which secured the first and second places in the constrained and unconstrained conditions respectively. We start off with a simple end-to-end approach with Transformers and then apply a gamut of ideas like replacing the encoder blocks with Conformers, encoder pre-training, etc., to drastically improve our dev BLEU score from 1.02 to 20.22. Through our results, we also quantitatively demonstrate how much of an impact each of our ideas bring forth and sincerely hope that some of these ideas might be useful for researchers and practitioners alike working on low-resource speech translation problems.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyothi, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pages 2819–2826.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al.

2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Arun Baby, Anju Leela Thomas, N. L. Nishanthi, and TTS Consortium. 2016. Resources for Indian languages. In *CBBLR – Community-Based Building of Language Resources*, pages 37–43, Brno, Czech Republic. Tribun EU.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799. IEEE.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? *arXiv preprint arXiv:2106.01045*.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal. 2008. Recent efforts in spoken language translation. *IEEE Signal Processing Magazine*, 25(3):80–88.

William Chan, Navdeep Jaitley, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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*.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.

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

Gaurav Kumar, Matt Post, Daniel Povey, and Sanjeev Khudanpur. 2014. Some insights from translating conversational telephone speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3231–3235.

Hang Le, Florentin Barbier, Ha Nguyen, Natalia Tomashenko, Salima Mdhaffar, Souhir Gahbiche, Bougares Fethi, Benjamin Lecouteux, Didier Schwab, and Yannick Estève. 2021. On-trac'systems for the iwslt 2021 low-resource speech translation and multilingual speech translation shared tasks. In *International Conference on Spoken Language Translation (IWSLT)*.

H. Ney. 1999. Speech translation: coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 1, pages 517–520 vol.1.

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

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*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Kishore Prahallad, E Naresh Kumar, Venkatesh Keri, S Rajendran, and Alan W Black. 2012. The iiit-h indic speech databases. In *Thirteenth annual conference of the international speech communication association*.

Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, Gautam Mantena, Bhargav Pulugundla, Peri Bhaskararao, Hema A Murthy, Simon King, Vasilis Karaiskos, and Alan W Black. 2013. The blizzard challenge 2013–indian language task. In *Blizzard challenge workshop*, volume 2013.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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

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.

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*.