Responses to Reviewers' Comments

We thank all the reviewers for the time and effort for the detailed analysis of the paper. We have addressed the concerns of all the reviewers with the highlighted changes in the manuscript. We have answered all the comments by the reviewers, along with the section numbers in the updated manuscript. (comments of the reviewers are highlighted)

Reviewer: Metareview

Q1 - Compare the model trained on the proposed corpus with another S2S system, e.g. SeamlessM4T (https://github.com/facebookresearch/seamless_communication). This will justify your claims on whether your mode is SOTA.

Reply:

> The model trained is compared with the results on the SeamlessM4T model. Please see Table 4.

Q2 - Benchmark the model on the FLEURS dataset.

Reply:

> The Fleurs dataset is benchmarked on the Indic-S2UT model, and the results are given in the Table 4.

Q3 - The from-English directions could be benchmarked in the task of speech-to-text translation.

Reply:

> The from-English direction data is benchmarked for the speech-to-text translation task. Please see Table 5.

Q4 - Report the language pairs of your experiments. For example, table 3 reports BLEU scores for into English translation.

> **Reply:** The language pairs are explicitly given with the results in the Tables 4 and 5.

Q5 - More information could be added on the languages. What is the original language of the programme? It appears that the dataset is not many-to-many, but many-to-one (English), unless all talks are dubbed into all 14 languages.

> **Reply:** All the 14 languages are original to the dataset, and the dataset is many-to-many (n-way parallel) as the entire speech data is dubbed in all the 14 languages, maintaining many-to-many property of the dataset. The same has been mentioned in the paper wherever required.

Q6 - What license will the data be released under?

Reply:

> The dataset will be released with CC-BY-4.0 license. The same has been mentioned in the Ethics section.

Q7 - Is the entire dataset validated or only the test set?

Reply:

> Yes, the entire dataset is validated and the detailed validation process is discussed in the Section 2.5

Reviewer: 1 (Reviewer Zqdi)

W1: Besides the general S2ST datasets, Indic datasets should also be mentioned as related work.

Reply:

> As we stated in the paper as well, FLEURS is the only S2ST dataset available for Indic languages (which is already mentioned). For a broader scenario, we mentioned some of the Indic datasets for other downstream tasks such as ASR, MT, TTS and ST. Please see sec 1 para 1.

W2: For speech processing, how was speaker diariziation done, e.g., which model was used, …? Is there any quality check to see whether information is lost after this step?

Reply:

> "Speaker diarization is applied using Pyannote 3.0 8 to remove unwanted background human noise (e.g., murmuring) and improve the clarity of the speech. We manually verified the speech quality for all the data across all 14 languages after diarization and confirmed that no information was lost." Please see sec 2.3 (speech preprocessing) and Please see sec 2.5 for detailed validation process..

W3: Line 130: "Non-English content in English text and English content in Non-English texts, are removed.": Was this done manually or automatically? If automatically, which tools did you use?

Reply:

> "As the data is sourced from a highly reliable source, there are not many foreign language characters in the texts. In non-English texts, there are only English characters as noise in the text, and in English texts, there are some Hindi characters as noise. Hence, we remove the unwanted English and Hindi characters from the text". Please see sec. 2.3 (text preprocessing).

W4: Line 136: "The texts are subsequently segmented sentence-wise for all the languages": how is this done? Shouldn't segmentation and alignment be done together? After reading the upcoming section on alignment, I interpreted that you segment only English, and align all other text to English? Is this correct?

Reply:

> We independently segment texts sentence-wise for all languages using the Indic-NLP library. We then align sentences from specific languages to English using BertAlign, with English serving as the anchor language for alignment. (Please see sec. 2.3 (text pre-preprocessing)).

W5: What are the translation directions of the results of your model in Table 3? Is it from/to English?

Reply:

Please see reply to Metareview Q.4.

W6: There is no baseline to compare the quality of the provided dataset and model with. It would be good to evaluate your pretrained model on some other Indic translation test sets, and compare it to existing models.

Reply:

Please see reply to Metareview Q.1.

C1: Line 077: "In E2E S2ST, no intermediate output is generated between the encoder and the decoder": I think it's not about intermediate output between any encoder or any decoder, but about intermediate output between models (ASR, MT, TTS).

Reply:

"Our statement refers to the fact that no intermediate output is generated, and no intermediate steps are required between the encoder and decoder." The same is updated in the paper.

C2: Line 152: It is not clear what you use in the end for text alignment. Is it "Bertalign"? Why does having topk similar sentences matter, and why does only "Bertalign" provide that but not the other methods?

Reply:

"We use BertAlign to align texts across various Indic languages. By leveraging top-k similar sentences, BertAlign prioritizes semantically closest matches rather than exact translations, making it effective even when perfect translations are unavailable. Based on a manual evaluation of aligners, BertAlign consistently provides the most accurate alignments." The same is updated in the paper (sec 2.4). For manual evaluation, please see sec 2.5.

C3: Line 217: Did you call the architecture you built your model upon "SOTA", or did you call your own model "SOTA"? If you call your own model "SOTA", you need to give qualitative scores comparison with existing models to show that your model is indeed the best.

Reply:
Please see reply to Metareview Q.1.

Reviewer: 2 (Reviewer Stn1)

*Incomplete benchmarking of the dataset *Incomplete comparison to related datasets

Reply:

Please see Table 4. We could not present results on the language pairs with Indic as target language since the unit-based HIFI-GAN vocoder is not yet available for Indic languages. In Table 4, we present results on the related dataset as well.

*Extend benchmarking with other modalities (MT) and baseline systems (NLLB?) *Further compare to recent related resources like 2M-Flores: https://huggingface.co/datasets/facebook/2M-Flores-ASL & https://arxiv.org/abs/2412.08274

Reply:

Please see Table 4 and 5 for results on SeamlessM4T. Yes, we can benchmark Indic-S2ST for other modalities as well. The sole purpose of this paper is to present dataset and benchmarking system for S2ST task as various datasets and models exist for other modalities but not for S2ST in Indic languages.Thus the question raised is outside the scope of the work presented.

We again thank all the reviewers for their time and effort for evaluating our work. We hope we are able to answer all the questions and concerns of the reviewers with the updated paper. We would like to answer the further questions, if any. Thanks again.