Indic-S2ST: a Multilingual and Multimodal Many-to-Many Indic Speech-to-Speech Translation Dataset

Nivedita Sethiya
IIT Indore
Simrol, Indore, MP
India 453552
phd2201201003@iiti.ac.in

Puneet Walia
GNDEC
Ludhiana, Punjab
India 141006
puneet2121091@gndec.ac.in

Chandresh Kumar Maurya
IIT Indore
Simrol, Indore, MP
India 453552
chandresh@iiti.ac.in

Abstract

Speech-to-Speech Translation (S2ST) converts speech from one language to speech in a different language. While various S2ST models exist, none adequately support Indic languages, primarily due to the lack of a suitable dataset. We fill this gap by introducing Indic-S2ST, a multilingual and multimodal many-to-many S2ST data of approximately 600 hours in 14 Indic languages, including Indian-accented English. To the best of our knowledge, this is the largest data for the S2ST task with parallel speech and text in 14 scheduled Indic languages. Our data also supports Automatic Speech Recognition (ASR), Text-to-Speech (TTS) synthesis, Speech-to-Text translation (ST), and Machine Translation (MT) due to parallel speech and text alignment. Thus, our data may be useful to train a model like Meta's SeamlessM4T for Indic languages. We also propose Indic-S2UT, a discrete unit-based S2ST model for Indic languages. To showcase the utility of the data, we present baseline results on the Indic-S2ST data using the Indic-S2UT. The dataset and codes are available at https://github.com/Nivedita5/Indic-S2ST/blob/main/README.md.

1 Introduction

As of September 2024, India's population stands at 1.43 billion, with Indic languages spoken by 17.78% of the world's population, underscoring their global significance.[1] This highlights the need to build a linguistic bridge to overcome communication barriers among speakers of Indic languages. Currently, many rely on foreign languages like English to communicate across language boundaries (Sethiya and Maurya, 2025). Therefore, a system capable of translating speech between Indic languages is warranted. Several S2ST datasets such as CVSS (Jia et al., 2022b), MaSS (Boito et al., 2019), SpeechMatrix (Duquenne et al., 2022), STC (Shimizu et al., 2014), Fleurs (Conneau et al., 2023), and Fisher (Post et al., 2013) are available, detailed in Table 1 (more details in (Gupta et al., 2024)). However, none of these datasets include Indic languages, except for Fleurs, which contains only 12 hours (as claimed by the authors) of read speech. Moreover, the data in Fleurs is neither validated nor fully aligned, and the authors (Conneau et al., 2023) acknowledge that some speech samples are missing (see sec. 2.8). However, there exist many Indic speech datasets for various tasks such as ASR (IndicVoices (Javed et al., 2024)), TTS (IndicVoices-R (Sankar et al., 2024)), MT (Samanantar (Ramesh et al., 2022)), and ST (IndicST (Sethiya et al., 2025)), to name a few. It is well known that models such as large language models (LLMs) achieve better performance on downstream tasks when trained on large volumes of high-quality data. Therefore, there is a significant need for a dedicated Indic speech-to-speech translation dataset.

DS	SS	TS	Hours	Lang	Indic
Fisher	Rd	Syn	127	1	✗
STC	Ip	Rl	31	1	✗
MaSS	Sp	Rl	20	8	✗
CVSS	Rd	Syn	181	21	✗
LibriS2S	Rd	Rl	52	1	✗
Speech-Matrix	Sp	Rl	1537	17	✗
Fleurs	Rd	Rl	12	102	✔
Indic-S2ST	Rd	Rl	42	14	✔

Table 1: Statistics of existing datasets for S2ST. (DS: Datasets, SS: Speech Synthesis, TS: Type of Speech, Rd: Read speech, Ip: Interpretation speech, Sp: Spontaneous speech, Syn: Synthetic speech, and Rl: Real speech). The details are sourced from the respective publications.

To solve the above problem, we introduce the

[1] https://www.worldometers.info/world-population/%20india-population/

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3766–3775

Indic-S2ST dataset that is larger than Fleurs and manually validated, thus ensuring high quality. Indic-S2ST data contains Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Manipuri (mni), Marathi (mr), Oriya (or), Punjabi (pa), Tamil (ta), Telugu (te), Urdu (ur), and Indian-accent English (en). The dataset is many-to-many (14-way parallel) as all the speech data is dubbed in all 14 languages. Throughout the paper, these languages are referred to by their respective ISO codes. Our contributions are as follows:

- Indic-S2ST, a multilingual, multimodal many-to-many S2ST data (mutliparallel data) of approximately a total of 600 hours (around 40 hours of speech per language), recorded in 14 Indic languages
- Aligned speech-text pairs for all 14 languages (making it a total of 196 language pairs for the S2ST task), making it useful for various tasks such as ASR, MT, TTS, ST, and language identification (LID)
- Indic-S2UT, a discrete unit-based speech-to-speech translation model is trained, and baselines for both S2ST task (Indic→English) and ST task (English→Indic, as unit-based vocoder is unavailable for Indic languages).[2]
- Comparative results for both the S2ST (Indic→English) and ST (English→Indic) tasks on SeamlessM4T model to act as a baseline.

1.1 Task Definition

S2ST tasks can be approached in two ways: End-to-End (E2E) or cascaded S2ST. In E2E S2ST, no intermediate output is generated between the encoder and the decoder, and no intermediate steps are required between the encoder and decoder. Cascaded S2ST, on the other hand, is implemented using a combination of ASR, MT, and TTS or through ST followed by TTS. The formal definition of the E2E S2ST task is as follows: Given the dataset $D = \{(\mathbf{S}, \mathbf{T})\}_{i=1}^{n}$, where $\mathbf{S} = \{s_1, s_2, .., s_x\}$ is the source language speech feature vector and $\mathbf{T} = \{t_1, t_2, .., t_y\}$ is the target language speech feature vector. The x and y are the lengths of the source and target speech in frames, respectively. The model is optimized wrt the negative log-likelihood $\sum_{i=1}^{n} -\log p(\mathbf{T}_i|\mathbf{S}_i;\theta)$ where the conditional probability is defined as:

$$p(\mathbf{T}|\mathbf{S};\theta) = \prod_{k=1}^{y} p(t_k|t_{<k}, \mathbf{S};\theta) \quad (1)$$

In the above equation, θ denotes the model parameters.

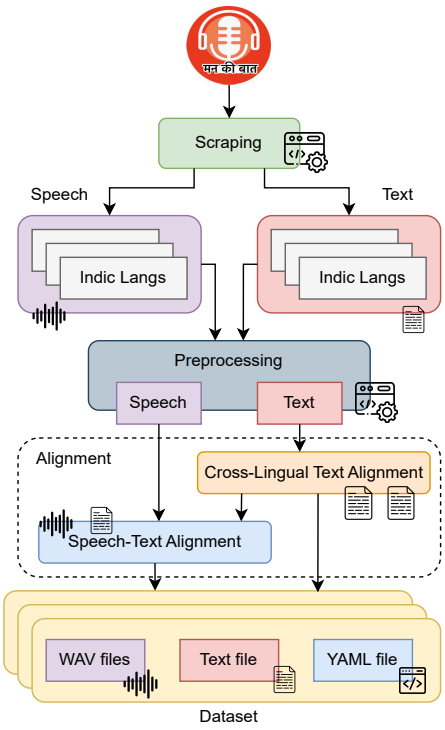

Figure 1: The flow diagram above illustrates the methodology used to curate the Indic-S2ST dataset, outlining all the processes involved. *Indic Langs* represents all the Indic languages included in the dataset, with each language following the same process in parallel.

2 Dataset Curation Methodology

This section contains the methodology used to curate the dataset for S2ST for Indic languages, including source details, preprocessing, alignment, and statistics, as depicted in Figure 1.

2.1 Data Source

We use the initiative by the Government of India, Mann Ki Baat as a source to collect the data.[3,4,5] In this program, the Prime Minister of India addresses countrymen monthly, which is broadcast

[2]Note: we do not present results of other possible tasks likely due to other datasets available for the same and space available here.

[3]https://www.narendramodi.in/mann-ki-baat
[4]https://soundcloud.com/narendramodi/sets
[5]https://www.narendramodi.in/mann-ki-baat

across the nation in multiple languages after dubbing by professional translators. The broadcast is accompanied by human-generated transcripts. We collect 90 talks starting from Oct 2014 in 14 different languages, including Indian-accented English. Each talk typically lasts around 30 minutes.[6]

2.2 Scraping

The speech recordings and corresponding transcripts needed for the dataset are available across various sources (mentioned in §2.1). We employ different web scraping tools, including Selenium, BeautifulSoup, and Soundfile, to extract the data from these sources.[7,8,9]

2.3 Preprocessing

The raw data contains noise and is not directly consumable by models. Hence, we apply data preprocessing techniques to clean the raw data.

Speech Preprocessing: The speech data extracted from the source is in raw audio format. The audio files are converted to WAV format at a sampling rate of 256kbps and a frequency of 16kHz. Additionally, speaker diarization is applied using Pyannote 3.0 to remove unwanted background human noise (e.g., murmuring), hence, improves the clarity of the speech and making the speech data more reliable.[10,11] We manually verified the speech quality for all the data across all 14 languages after diarization and confirmed that no information was lost (see 2.5 for detailed validation process).

Text Preprocessing: Inconsistent quotation marks are removed to standardize the transcripts into the texts. 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 (Latin) characters as noise in the text, and in English texts, there are some Hindi (Devnagari) characters as noise. Hence, we remove the unwanted English (Latin) and Hindi (Devnagiri) characters from the text.[12] Patterns like date, and header/footer are eliminated to ensure data cleanliness. We remove only the characters that are noise to the data, and keep the sentences. Blank lines and lines corresponding to audio files durations shorter than 0.025 seconds, which offer minimal usable information, are also removed. Any audio file that is shorter than 0.025 second, will have a maximum of 2 words and will not constitute any meaning, hence will be irrelevant for the dataset. 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.

2.4 Alignment

Alignment is the process of finding the relationship between words in two different languages or modalities. The raw data is aligned at two levels: (a) aligning texts in two different languages, and (b) aligning speech and text of the same language.

Cross-Lingual Text Alignment: It aligns texts from two different languages based on semantic or contextual similarities, resulting in a sentence-by-sentence correspondence between text files of two different languages. Various techniques are available for cross-lingual text alignment, such as Bleualign, Hunalign (Varga et al., 2008), Vecalign (Thompson and Koehn, 2019), etc.[13] Bertalign (Liu and Zhu, 2023) is the most suitable method, as it identifies the top-k most semantically similar target sentences. Hence, 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. Multilingual alignment of all the Indic languages of the dataset is performed with English as the anchor language, and hence the sentences are aligned. We align all segmented sentences across the languages, with the English sentences serving as the alignment point.

Speech-Text Alignment: It aligns speech frames with the corresponding words of the text of the same language, a technique also known as *Forced Alignment*. While several forced aligners, such as MFA (McAuliffe et al., 2017) and Prosodylab-Aligner (Gorman et al., 2011), are

[6] All the permissions required to use the data have been obtained.

[7] https://www.selenium.dev/

[8] https://beautiful-soup-4.readthedocs.io/en/latest/

[9] https://pypi.org/project/soundfile/

[10] https://huggingface.co/pyannote/speaker-diarization-3.0

[11] The background noise is the unwanted human voices in the audio, removed using the speaker segmentation technique.

[12] https://github.com/anoopkunchukuttan/indic_nlp_library

[13] https://github.com/rsennrich/Bleualign

Langs	Audio files	Speech hours	Audio file duration			Sent	Tokens	Utter	Sent Length	
			Max	Min	Avg				Max	Avg
As	17,297	37.78	239.44	0.04	7.63	17,297	250,930	46,536	268	15.04
Bn	17,297	43.45	1051.24	0.16	8.70	17,297	256,890	43,444	287	15.37
En	17,297	37.22	252.58	0.03	7.60	17,297	327,743	41,956	315	18.45
Gu	17,297	42.94	438.24	0.24	8.56	17,297	285,796	44,942	310	17.03
Hi	17,297	46.34	552.40	0.04	9.28	17,297	344,078	31,598	372	20.31
Kn	17,297	45.91	694.72	0.04	9.14	17,297	210,226	59,498	202	12.82
Ml	17,297	40.90	789.72	0.04	8.13	17,297	199,224	65,123	199	12.21
Mni	17,297	37.65	1899.08	0.04	8.14	17,297	323,312	38,733	258	19.32
Mr	17,297	46.72	958.64	0.08	9.29	17,297	250,445	52,459	266	15.05
Or	17,297	40.52	109.16	0.04	8.12	17,297	278,479	34,551	276	16.65
Pa	17,297	43.63	306.60	0.12	8.69	17,297	354,034	30,391	393	20.77
Ta	17,297	48.63	404.48	0.16	9.64	17,297	219,570	60,555	210	13.38
Te	17,297	42.06	191.68	0.04	8.38	17,297	217,382	56,715	193	13.23
Ur	17,297	40.13	251.80	0.04	8.09	17,297	352,247	29,876	393	20.91
Total	242,158	598.88	1899.08	0.03	8.53	242158	3,870,356	636,377	393	16.46

Table 2: Statistics of IndicS2ST dataset. All the languages are denoted by their ISO codes. Audio file durations are in ms and sentence length is in the number of tokens.

available, they do not produce satisfactory results for Indic languages on manual validation. To overcome this, we use different aligners for different languages: Gentle for English, Aeneas for Bn, Gu, Hi, Kn, Ml, Mr, Or, Pa, Ta, Te, and Ur, and MMS (Pratap et al., 2024) for As and Mni.[14,15] Using the forced aligners, we segment the WAV files for all languages based on sentence-level timestamps on text files obtained after cross-lingual text alignment.

For all 14 Indic languages, we compile the text files (denoted as $.lang$, where $lang$ is the ISO code). A TSV and YAML file is then generated for all languages, mapping each sentence in the text file to its corresponding segmented WAV file. The details of human validation for the dataset is given in the subsequent section. The combined processed data of all languages, thus curated is termed as **Indic-S2ST**.

2.5 Human Validation

To ensure the validity of the Indic-S2ST dataset, we employ human evaluators for all language pairs after each data processing step: speech preprocessing, text preprocessing, cross-lingual text alignment, and speech-text alignment. A total of 26 undergraduate and postgraduate students, both male and female, aged 18 to 35, participated in the validation process. For each language pair, we assign two experts per pair, proficient in both script and speech, with fluency in English and the respective Indic language. We employed a human validation process using a 5-point scale, where each speech-text pair was evaluated by 28 annotators. The scoring criteria are defined as follows: 0: completely noisy, 1: either speech or text not clear, 2: marginal noise in the data, 3: acceptable quality with minimal errors, 4: data with minimal misalignments, and 5: perfect quality data. We retained only speech-text pairs that received a score of ≥ 3, based on the following rationale: scores of 3 or higher consistently reflected audible speech and well aligned text quality, while lower scores indicated major errors in the data. Audio clips and sentences with alignment scores of 1 or 2 are discarded. Additionally, we checked the scores with both the human evaluators employed for the specific language pair. The entire dataset (both speech and text) is validated manually. Instances receiving a score of 3 undergo further review before a final decision is made on retention or removal. This process ensures consistency across all language pairs, maintaining the dataset's n-way parallel structure.

2.6 Data Statistics

Table 2 provides statistics of the Indic-S2ST dataset curated in the previous section. The data for each language spans 90 Mann Ki Baat talks with nearly

[14] https://github.com/lowerquality/gentle
[15] https://github.com/readbeyond/aeneas

Langs	Speech Hours		No. of Sentences	
	Indic-S2ST	Fleurs	Indic-S2ST	Fleurs
As	37.78	6.25	17,297	1961
Bn	43.45	6.01	17,297	1981
En	37.22	4.64	17,297	1938
Gu	42.94	4.86	17,297	1996
Hi	46.34	4.82	17,297	1702
Kn	45.91	5.41	17,297	1798
Ml	40.90	5.67	17,297	1955
Mni	37.65	-	17,297	-
Mr	46.72	6.21	17,297	1992
Or	40.52	3.30	17,297	1327
Pa	43.63	5.13	17,297	1588
Ta	48.63	5.88	17,297	1886
Te	42.06	5.83	17,297	1757
Ur	40.13	4.64	17,297	1588

Table 3: Comparison of speech hours and number of sentences of Indic-S2ST and Fleurs. Fleurs does not support the Manipuri language. An equal no. of sentences of Indic-S2ST represents a 14-way parallel (many-to-many) dataset, which Fleurs lack.

17,297 segmented audio files with parallel sentences/texts. The table presents key statistics for both speech and text. For speech, statistics for each language include the number of audio files, the speech hours, the number of utterances, and audio file duration (max, min, and avg). For text, the statistics include the number of sentences, the number of tokens, and the sentence length(max and avg). As the dataset is many-to-many, the data distribution across all the languages is the same in terms of the number of audio files and the number of sentences.[16]

2.7 Data utility

Since the Indic-S2ST data has parallel speech-text pairs for all languages, a variety of tasks can be performed on the data. For example, ASR (Javed et al., 2024) (Chadha et al., 2022), MT (Bala Das et al., 2024) (Dixit et al., 2023), TTS (Sankar et al., 2024) (Prakash and Murthy, 2022), and LID (Javed et al., 2023) models for Indic languages can be trained/fine-tuned. Besides that, ST (Sethiya et al., 2024) (Khurana et al., 2024) and S2ST (Jia et al., 2022a) models can be developed. Also, linguistic and morphological analyses can be done on Indic languages using Indic-S2ST. In this work, we showcase the utility of the Indic-S2ST data on the S2ST task, which is the most difficult among all of these tasks (due to cross-lingual, multimodality, acoustic ambiguity, etc.), in the next section.

2.8 Data comparison: Indic-S2ST & Fleurs

Table 3 presents the total speech hours and number of sentences in Indic-S2ST and FLEURS (as of 10 February 2025) to highlight the necessity of Indic-S2ST.[17] The table reveals that the number of sentences in FLEURS is not n-way parallel, varying across languages, despite claims to the contrary by the authors (Conneau et al., 2023). Additionally, the dataset lacks cleanliness and manual validation, making it challenging to train models and achieve comparable results. In contrast, Indic-S2ST provides n-way parallel data across all languages, undergoes manual validation, and facilitates model training. Also, Fleurs have American-accent English and Pakistani-accent Urdu, while Indic-S2ST have Indian-accent English and Indian-accent Urdu.

3 Experiments & Results

3.1 Model

We adopt an end-to-end (E2E) speech-to-speech translation (S2ST) approach to evaluate the Indic-S2ST dataset, motivated by the reduced error propagation typically associated with E2E systems compared to cascaded architectures (Gupta et al., 2024). We propose the Indic-S2UT architecture that specifically leverages the state-of-the-art S2UT model (Lee et al., 2021), which is a sequence-to-sequence framework that operates on discrete units for speech modeling.

The S2UT architecture comprises a transformer-based speech encoder and a decoder that predicts sequences of discrete acoustic units. The encoder is enhanced with auxiliary learning objectives to improve representation learning, and employs connectionist temporal classification (CTC) decoding to align input speech with target unit sequences. For waveform reconstruction, a pre-trained vocoder is used to synthesize speech from the predicted discrete units.

Discrete units are extracted using the HuBERT base model (Hsu et al., 2021), where representations from the 6th transformer layer are clustered using k-means (Lakhotia et al., 2021) to obtain 100 quantized units. This discretization process

[16] A sentence can have multiple utterances. An utterance is one of many ways to saying the same sentence.

[17] https://huggingface.co/datasets/google/fleurs/tree/main/data

Language Pairs	Indic-S2UT				SeamlessM4T	
	Indic-S2ST		Fleurs			
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
As→En	**21.99**	**75.88**	18.49	56.44	15.74	36.34
Bn→En	**23.04**	**80.65**	19.46	60.43	15.72	35.97
Gu→En	22.47	58.46	20.33	63.65	**23.07**	**66.14**
Hi→En	**26.08**	**85.88**	24.87	80.67	23.55	40.30
Kn→En	20.26	69.53	**21.05**	**72.06**	17.66	34.45
Ml→En	**18.85**	57.22	18.21	**57.44**	15.66	33.25
Mni→En	**18.77**	**65.49**	-	-	9.29	35.08
Mr→En	**15.29**	**53.91**	13.98	49.03	14.42	34.59
Or→En	**8.55**	**47.20**	4.01	43.76	0.08	19.44
Pa→En	19.35	70.46	**29.18**	**75.39**	24.74	34.16
Ta→En	**21.56**	**79.58**	13.07	68.44	9.81	38.03
Te→En	**20.24**	**78.43**	17.79	67.91	15.97	37.07
Ur→En	4.52	39.34	**9.21**	**44.80**	7.42	18.60

Table 4: S2ST task performance evaluation of Indic-S2UT pre-trained on the Indic-S2ST and Fleurs dataset and a a pre-trained SeamlessM4T results with translation direction from Indic speech → English speech. We report the normalized BLEU and chrF++ score here. The test set is the same for all the models. Best BLEU and chrF++ scores are highlighted for all the language pairs.

enables the model to learn unit-level representations of speech. For textual targets, CTC decoding with subword tokenization is employed to manage sequence length effectively to align input speech with target units generated by the HuBERT model.

To generate natural-sounding speech, we utilize a pre-trained unit-based HiFi-GAN vocoder (Kong et al., 2020), which incorporates duration prediction to improve alignment and rhythmicity of the synthesized output. For Indic language speech, discrete units are extracted using the multilingual and robust MR-HuBERT model (Shi et al., 2023), which provides better coverage across diverse linguistic features. This model is trained with the translation direction from the Indic speech to English speech. The S2UT model is implemented using the fairseq toolkit (Wang et al., 2020), with model configurations defined through YAML files that are publicly released alongside the Indic-S2ST dataset.

As a unit-based HiFi-GAN vocoder is currently unavailable for generating target Indic speech, and decoding intermediate discrete units is necessary for output generation, we adopt a unit-based text decoder based on the SpeechUT model (Zhang et al., 2022), which predicts the target Indic text (instead of the target Indic Speech). This allows us to demonstrate the usability of the Indic-S2ST dataset for Indic speech through the ST task (target text and not speech). This decoder model adopts the transformer architecture (Vaswani, 2017), consisting of a text embedding layer, multiple stacked transformer decoder layers, and a final output projection layer. It autoregressively generates the target text sequence from left to right, conditioned on the discrete unit representations produced by the unit encoder. We train this ST model to translate from English speech to text in Indic languages.

Training Details: We split the dataset in a ratio of 70:20:10 for train/test/dev sets. The model is optimized with the label smoothing loss. We train the models for 400k steps using the Adam optimizer with a learning rate of 0.0005 and a dropout rate of 0.1. We apply an inverse square root learning rate decay schedule, and each training step processes up to 20,000 tokens per batch. We trained the Indic-S2UT model on the Indic-S2ST dataset and the Fleurs dataset for a better comparison.

All the pre-processing and curation tasks for the Indic-S2ST dataset and the training and inference of the S2ST and ST tasks on the Indic-S2UT model are executed on NVIDIA A100-SXM4 with a VRAM of 40GB.

SeamlessM4T: The SeamlessM4T architecture (Barrault et al., 2023) builds on the UnitY framework (Inaguma et al., 2022), enabling joint optimization across modalities. Its text encoder and

Language Pairs	Indic-S2UT BLEU	Indic-S2UT chrF++	SeamlessM4T BLEU	SeamlessM4T chrF++
En→As	**14.91**	**42.09**	6.83	39.72
En→Bn	**18.35**	**51.89**	14.04	41.54
En→Gu	**22.73**	**59.36**	16.64	52.85
En→Hi	**29.82**	**55.61**	27.89	51.76
En→Kn	**17.61**	**62.63**	11.36	54.39
En→Ml	**18.25**	**55.57**	8.87	38.96
En→Mni	**12.94**	**40.90**	1.09	8.78
En→Mr	**18.27**	**52.46**	13.23	40.56
En→Or	**18.24**	**58.80**	12.52	50.22
En→Pa	20.47	51.67	**21.92**	**56.62**
En→Ta	**24.45**	**57.44**	11.14	52.89
En→Te	**17.60**	**67.07**	15.74	57.24
En→Ur	17.52	48.75	**21.04**	**52.88**

Table 5: ST task performance evaluation of Indic-S2UT pre-trained on the Indic-S2ST dataset and SeamlessM4T model with translation direction from English speech → Indic text. We report the normalized BLEU and chrF++ score here. The test set is the same for all the models. Best BLEU and chrF++ scores are highlighted for all the language pairs.

decoder are initialized from the NLLB translation model (Team et al., 2022), while speech inputs are processed using an enhanced Wav2Vec-BERT 2.0 encoder (Chung et al., 2021) with additional codebooks. A modality adapter (Zhao et al., 2022) aligns speech with text representations, and a text-to-unit (T2U) module generates discrete speech units, which are converted to audio using a HiFi-GAN vocoder (Kong et al., 2020). We chose not to fine-tune SeamlessM4T on the Indic-S2ST dataset, since fine-tuning would mask the independent contribution of the dataset itself. Rather than adapting SeamlessM4T to the same data, a direct comparison of its results with those of the Indic-S2UT model more clearly highlights the relevance and effectiveness of the Indic-S2ST dataset.

Evaluation Details: We evaluate the Indic-S2UT (pretrained on both the Indic-S2ST and Fleurs datasets individually) and Seamless models on the same test sets. For inference on the Indic-S2ST and FLEURS datasets using the Indic-S2UT model for the S2ST task, we first generate unit sequences using beam search with a beam width of 10. These unit sequences are then converted into speech using a pre-trained unit-based HiFi-GAN vocoder. As no standardized metric directly evaluates S2ST outputs, we assess translation quality by automatically transcribing the generated speech and comparing it to the reference text. The test set remains the same for all the model evaluations for the S2ST and ST tasks, mentioned in the training details of §3.1. The test set has some of the overlapping speakers (some of the voices are the same), but the speech content of those speakers in the training data is different (no overlap in the train and test set).

Specifically, we employ a wav2vec 2.0 ASR model (Vaessen and Van Leeuwen, 2022), which achieves a word error rate of 1.9% on the LibriSpeech dataset (Panayotov et al., 2015), to transcribe the synthesized speech into text. BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) scores are then computed between the ASR outputs and the ground-truth translations. To ensure a fair comparison, we follow the same evaluation protocol for the SeamlessM4T model by transcribing its generated speech using the same ASR system, and also the test set remains the same for generating all the results. Given the unavailability of a unit-based HiFi-GAN vocoder for Indic languages, we additionally evaluate the utility of the parallel Indic speech data through the ST task with the reverse direction, i.e., from English speech to Indic text translations. Here, the intermediate units are decoded into text using the SpeechUT decoder, and performance is again measured via BLEU and chrF++ scores. SeamlessM4T, a state-of-the-art LLM-based model for S2ST, is used as a baseline for both S2ST and ST evaluations.

3.2 Results

Table 4 presents the benchmark results of the Indic-S2UT model on the Indic-S2ST and FLEURS datasets for S2ST task across all Indic-to-English language pairs, along with the baseline results from the SeamlessM4T model. The Indic-S2UT model achieves strong and consistent performance across most languages. The highest BLEU and chrF++ scores are observed for Hindi, likely due to the larger amount of Hindi data used during the pre-training of the HuBERT Base model. In contrast, lower scores for Urdu and Punjabi may result from the limited availability and lower acoustic quality of the data used to train HuBERT for these languages, as well as their distinctive phonetic characteristics.

When comparing results across datasets, Indic-S2UT demonstrates clear improvements on the Indic-S2ST dataset compared to the FLEURS dataset. This highlights both the effectiveness and

domain alignment of Indic-S2ST for evaluating real-world speech-to-speech translation tasks.

Table 5 summarizes the benchmark results of the Indic-S2UT model on the Indic-S2ST dataset for ST task, translating English speech into Indic text. The Indic-S2UT model consistently outperforms the SeamlessM4T baseline across all evaluated languages, confirming its robustness and adaptability to Indic data.

It is important to emphasize that this comparison primarily aims to validate the reliability and relevance of the Indic-S2ST dataset, rather than to establish model superiority. The Indic-S2UT model is trained on a smaller but cleaner and more carefully curated set of real-world data, whereas SeamlessM4T is trained on large-scale, web-mined speech data, which is often noisy and may include freely available Mann Ki Baat recordings. Despite SeamlessM4T's exposure to similar data, Indic-S2UT achieves higher accuracy, demonstrating that compact, high-quality, and domain-specific data can yield better results than large, noisy datasets.

Overall, these findings highlight the Indic-S2ST dataset as a reliable and representative benchmark for Indic speech-to-speech translation. Given the scarcity of real-voice parallel data for Indic languages, the dataset provides an essential foundation for advancing research in multilingual and low-resource S2ST.

4 Conclusion and Future Works

In the present paper, Indic-S2ST, a multimodal multilingual many-to-many Indic S2ST dataset is proposed. The dataset contains parallel speech and text for 14 Indic languages. We also present Indic-S2UT model pretrained for the S2ST task on the Indic-S2ST for the Indic→English speech pairs and for the ST task for the English→Indic speech-text pairs. We also present baseline results on the SeamlessM4T model for both the S2ST and ST tasks on respective language pairs. Though initial results are plausible, there is still room for improvement. In the future, we plan to provide n-way speech-to-speech translation results for all the Indic languages present in the Indic-S2ST dataset. Further, the present work might open opportunities for researchers to work on Indic speech-to-speech translation.

5 Ethics

We do not foresee any ethical risks as the data is already used for many research purposes and is publicly available. We recognize the responsibility of releasing a dataset that holds great significance for Indic language speakers. While we have carefully processed and curated the dataset, the content obtained from the original source remains unchanged. The ideas and opinions reflected in the data are entirely those of the source, and we do not assume responsibility for any inaccuracies. The authors also do not endorse any form of bias related to gender, religion, caste, faith, or actions that could harm the sentiments of any living being. The dataset will be released under the CC-BY-4.0 license.

Limitations & Ethics

While Indic-S2ST includes a handful number of Indic languages, several important languages are still missing from the dataset and should be incorporated. The dataset features real human voices, which is crucial, but each language is represented by a limited number of speakers. Also, the data is sourced from a single genre of formal-broadcasted speeches. The evaluation set has the same voices, but the speech content in the training data is different, and also there is no overlap in the train and test set. Additionally, the number of hours of speech data is limited and should be expanded to achieve better results. While this paper presents results of only 13 language pairs for speech to speech translation, 183 language pairs still remain to be benchmarked on the Indic-S2ST dataset.

References

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, Bidyut Kr. Patra, and Asif Ekbal. 2024. Multilingual neural machine translation for indic to indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(5):1–32.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Marcely Zanon Boito, William N Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2019. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *arXiv preprint arXiv:1907.12895*.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: Asr toolkit for low resource indic languages. *arXiv preprint arXiv:2203.16512*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. 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)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and 1 others. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian acoustics*, 39(3):192–193.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2024. Direct speech-to-speech neural machine translation: A survey. *arXiv preprint arXiv:2411.14453*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2022. Unity: Two-pass direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2212.08055*.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12942–12950.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, and 1 others. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *arXiv preprint arXiv:2403.01926*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. Cvss corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*.

Sameer Khurana, Chiori Hori, Antoine Laurent, Gordon Wichern, and Jonathan Le Roux. 2024. Zerost: Zero-shot speech translation. In *Interspeech 2024*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, and 1 others. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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*.

Anusha Prakash and Hema A Murthy. 2022. Exploring the role of language families for building indic speech synthesisers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:734–747.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and 1 others. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *arXiv preprint arXiv:2409.05356*.

Nivedita Sethiya and Chandresh Kumar Maurya. 2025. End-to-end speech-to-text translation: A survey. *Computer Speech & Language*, 90:101751.

Nivedita Sethiya, Saanvi Nair, and Chandresh Maurya. 2024. 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)*, pages 9019–9024.

Nivedita Sethiya, Saanvi Nair, Puneet Walia, and Chandresh Maurya. 2025. Indic-st: A large-scale multilingual corpus for low-resource speech-to-text translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(6):1–25.

Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. 2023. Multi-resolution hubert: Multi-resolution speech self-supervised learning with masked unit prediction. *arXiv preprint arXiv:2310.02720*.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *LREC*, pages 670–673.

NLLB Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348.

Nik Vaessen and David A Van Leeuwen. 2022. Fine-tuning wav2vec2 for speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7967–7971. IEEE.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2008. Parallel corpora for medium density languages. In *Recent advances in natural language processing IV: selected papers from RANLP 2005*, pages 247–258. John Benjamins Publishing Company.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. *arXiv preprint arXiv:2210.03730*.

Jinming Zhao, Hao Yang, Ehsan Shareghi, and Gholamreza Haffari. 2022. M-adapter: Modality adaptation for end-to-end speech-to-text translation. *arXiv preprint arXiv:2207.00952*.