UrHiOdSynth: A Multilingual Synthetic Corpus for Speech-to-Speech Translation in Low-Resource Indic Languages

Anonymous ACL submission

Abstract

Speech-to-Speech Translation (S2ST) focuses on generating spoken output in a target language directly from spoken input in a source language. Despite progress in S2ST modeling, low-resource Indic languages remain poorly supported, primarily because large-scale parallel speech corpora are unavailable. We present UrHiOdSynth, a three-language parallel S2ST dataset containing approximately 75 hours of speech across Urdu, Hindi, and Odia. The corpus consists of 10,735 aligned sentence triplets, with an average utterance length of 8.45 seconds. To our knowledge, UrHiOdSynth represents the largest multi-domain resource offering aligned speech and text for S2ST in this language context. Beyond speech-to-speech translation, the dataset supports tasks such as automatic speech recognition, speech-to-text translation, text-to-speech synthesis, and machine translation. This flexibility enables the training of unified multilingual models, particularly for low-resource Indic languages. The dataset and code are publicly available at https://github.com/UrHiOdsynth/UrHiOdsynth.

1 Introduction

Speech-to-speech translation (S2ST) aims to reduce language barriers by enabling direct communication between speakers of different languages (Gupta et al., 2025b). Recent advances in deep learning drive significant improvements in S2ST systems (Gong et al., 2025; Fang et al., 2023). Despite this progress, most existing approaches rely on a cascaded pipeline that combines automatic speech recognition, text-based machine translation, and text-to-speech synthesis (Wu et al., 2024). These cascaded systems perform well in practice because they leverage the availability of large-scale parallel text resources.

Although cascaded S2ST systems work well in many settings, they come with clear limitations. Because translation is carried out through multiple stages, errors introduced early in the pipeline tend to propagate to later components. In addition, converting speech into text makes it difficult to retain speech-specific information such as prosody, emotion, and cultural nuances, which are often diluted or lost in intermediate textual representations (Smith et al., 2022). To address these issues, recent research has shifted toward end-to-end S2ST models that directly translate speech from a source language into speech in a target language without relying on intermediate text (Jones et al., 2022; Sarim et al., 2025). While direct S2ST approaches promise more natural and expressive translations, they face significant challenges, most notably the scarcity of high-quality parallel speech datasets for many language pairs. Nevertheless, ongoing advancements in neural architectures and data generation strategies continue to push the boundaries of S2ST research.

Although models such as Translatotron (Nachmani et al., 2024) and SeamlessM4T (Barrault et al., 2023) enable direct speech-to-speech translation, they offer limited coverage of Indic languages. UrHiOdSynth addresses this gap by providing the first trilingual multi-domain synthetic parallel speech dataset for Hindi, Urdu, and Odia, as shown in Table 1.

Direct speech translation systems for Indic languages are severely constrained by the scarcity of parallel speech corpora. By employing synthetic data generation methods to construct UrHiOdSynth, we offer a scalable and practical solution for advancing direct S2ST research in low-resource Indic languages. The main contributions of our study are as follows:

- Creation of UrHiOdSynth, a high-quality synthetic tri-parallel speech and text dataset spanning Hindi, Urdu, and Odia for low-resource multilingual speech-to-speech translation.

- Development of benchmark baseline model for direct speech-to-speech translation on Hindi–Urdu, Hindi–Odia, and Urdu–Odia language pairs, establishing UrHiOdSynth as a foundational dataset for Indic S2ST research.

Language	# Sentences	# Hours	Mean (s)
Hindi	10735	23.27	7.80
Urdu	10735	27.66	9.27
Odia	10735	24.70	8.28
Total	**32205**	**75.63**	**8.45**

Table 1: Statistics of the UrHiOdSynth dataset across Hindi, Urdu, and Odia, reporting the number of sentences, total speech duration (in hours), and the mean duration (in seconds)

2 Related Works

The creation of large-scale natural parallel speech corpora is both time-consuming and costly, motivating researchers to explore alternative strategies for constructing and leveraging datasets for speech-to-speech translation (S2ST). Prior work has demonstrated the effectiveness of synthetic and multilingual speech resources through efforts such as Translatotron, VoxPopuli, and MuST-C. However, these datasets provide little to no coverage of Indic languages. Although this gap has recently been partially addressed by IndicSynth (Sharma et al., 2025) a multilingual synthetic speech dataset designed for audio deepfake detection and anti-spoofing across 12 low-resource Indic languages parallel speech resources for direct speech translation among Indic languages such as Hindi, Urdu, and Odia remain extremely limited.

Several large-scale multilingual speech datasets have been proposed for S2ST and related tasks. CVSS (Jia et al., 2022) is a massively multilingual speech-to-speech translation corpus containing sentence-level parallel S2ST pairs from 21 languages into English. FLEURS (Conneau et al., 2023) is a multilingual speech dataset intended to evaluate few-shot learning for universal speech representations, covering 102 languages with approximately 12 hours of supervised speech data per language. FLEURS extends the FLoRes-101 (Goyal et al., 2022) machine translation benchmark by providing speech data as an n-way parallel resource. FLoRes-101 itself is a high-quality multilingual benchmark consisting of 3,001 professionally translated Wikipedia sentences across 101 languages, enabling many-to-many evaluation over 10,100 translation directions with human verification to ensure reliability.

Additional multilingual speech corpora include MaSS (Boito et al., 2019), which was constructed using the CMU Wilderness dataset and provides approximately 20 hours of aligned speech across eight languages for both speech-to-text and speech-to-speech tasks, with quality validated by native speakers on a subset of the data. SpeechMatrix (Duquenne et al., 2022) is a large-scale multilingual S2ST dataset compiled from European Parliament recordings and represents one of the largest resources of its kind, offering approximately 418,000 hours of aligned speech across 136 language pairs. By minimizing the cosine loss in relation to LASER's multilingual text embedding space (Duquenne et al., 2021) proposed a method for mapping speech into fixed-dimensional embeddings. By leveraging shared latent representations, the technique enables direct speech-to-speech translation as well as cross-modal alignment between speech and text.

The pivot-language translation has been widely used to overcome the scarcity of direct parallel corpora. Several studies extract or generate bilingual data by translating through a high-resource pivot, such as using English to derive Persian–Italian pairs from non-parallel corpora (Ansari et al., 2017), or multiple pivot languages to enhance sparse corpora for Indic/East Asian translation (Dabre et al., 2015). Recent work employs pivot languages such as Hindi to bootstrap translation between low-resource language pairs, demonstrating improvements in synthetic corpus construction (Talwar and Laasri, 2025). The multilingual efforts, such as the Samanantar corpus, also exploit pivot-based extraction from English to expand data coverage across many Indic languages (Ramesh et al., 2022). Additionally, pivot translation methods have been examined in zero-resource settings using multilingual pretrained models (Imamura et al., 2023).

In recent years, several large-scale Indic speech corpora such as Nirantar (Javed et al., 2025), MahaDhwani (Bhogale et al., 2025), Svarah (Javed et al., 2023), Shrutilipi (Bhogale et al., 2023), and Dhwani (Javed et al., 2022) have been released. Although these resources collectively span multiple Indic languages and differ in both linguistic diversity and recording hours, none provide parallel speech pairs across any two Indic languages.

Dataset	Domain	Speech Nature	#Lang.	Hours	Hi–Ur–Od	#Uttr
Gupta et al. (Gupta et al., 2025a)	TED Talks	Synthetic	2	116	✗	121.4K
Mondal et al. (Mondal et al., 2024)	Audiobook	Spont./Synthetic	2	2	✗	✗
FLEURS (Conneau et al., 2023)	Wikipedia	Read	102	1.4K	✓	2.7K
SpeechMatrix (Duquenne et al., 2022)	EuroParl	Spontaneous	17	418K	✗	✗
CVSS (Jia et al., 2022)	Open-domain	Read/Synthetic	21	719	✗	✗
VoxPopuli (Wang et al., 2021)	EuroParl	Spontaneous	15	17.3K	✗	✗
LibriVoxDeEn (Beilharz et al., 2019)	Audiobook	Read	2	52.5	✗	✗
MASS (Boito et al., 2019)	Bible	Read	15	150	✗	✗
UrHiOdSynth (Ours)	Multi-domain	Synthetic	3	**75.63**[1]	✓	**32.2K**

Table 2: Comparative overview of existing speech-to-speech translation (S2ST) datasets and the proposed UrHiOdSynth corpus across domain, speech characteristics, language coverage, and scale. #Lang. denotes the number of languages and #Uttr is the number of utterances.

Consequently, they offer limited utility for training or evaluating direct S2ST models.

To bridge this gap, the present study introduces a new large-scale Hindi–Urdu-Odia parallel S2ST dataset specifically designed for low-resource conditions better aligned with the requirements of end-to-end speech-to-speech translation research for Indic languages.

3 Motivation

The UrHiOdSynth corpus is motivated by the growing need for robust multilingual speech technologies in increasingly interconnected linguistic communities. While substantial textual resources exist for many Indic languages (Ramesh et al., 2022), high-quality parallel speech data with domain coverage remains severely limited, constraining the development of natural and expressive speech systems. Spoken language encodes prosody, intonation, and emotion that cannot be captured by text alone, making speech-level supervision essential for advanced applications such as direct speech-to-speech translation and voice-based assistants (Sarim et al., 2025). By leveraging parallel text resources to construct a tri-lingual Hindi–Urdu–Odia speech corpus, UrHiOdSynth addresses language-specific phonetic and dialectal challenges while enabling inclusive, accessible, and education-oriented speech technologies.

4 Corpus Creation Pipeline

The main procedures for getting the UrHiOdSynth dataset ready are covered in this section.

4.1 Extracting Text pairs

Parallel Hindi–Urdu sentence pairs were curated from open-domain resources, specifically the CoRIL parallel corpus (Bhattacharjee et al., 2025). Parallel Odia text was generated by translating the Hindi sentences using the Google Translate API (Limbu, 2020). The trilingual parallel text is organized across three domains, namely government, general, and healthcare, to explicitly capture variation in topic, register, and sentence structure. This domain stratification expands linguistic coverage and supports systematic evaluation of domain robustness in speech translation models trained on the corpus. The corpus creation process follows the pipeline shown in Figure 1.

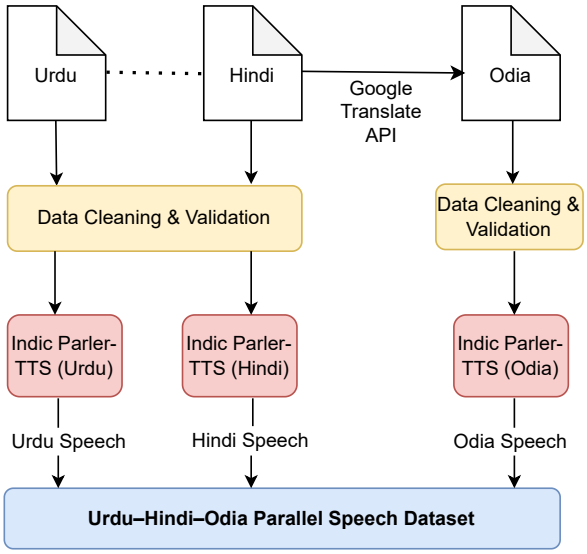

Figure 1: Pipeline for creating a trilingual Hindi–Urdu-Odia parallel speech corpus.

4.2 Text Preproccessing

To facilitate downstream data processing, we first apply text normalization, as the parallel text pairs were collected from online sources. Character representations are standardized using NFKC normalization (Hou et al., 2025). We then remove extraneous elements such as punctuation, escape charac-

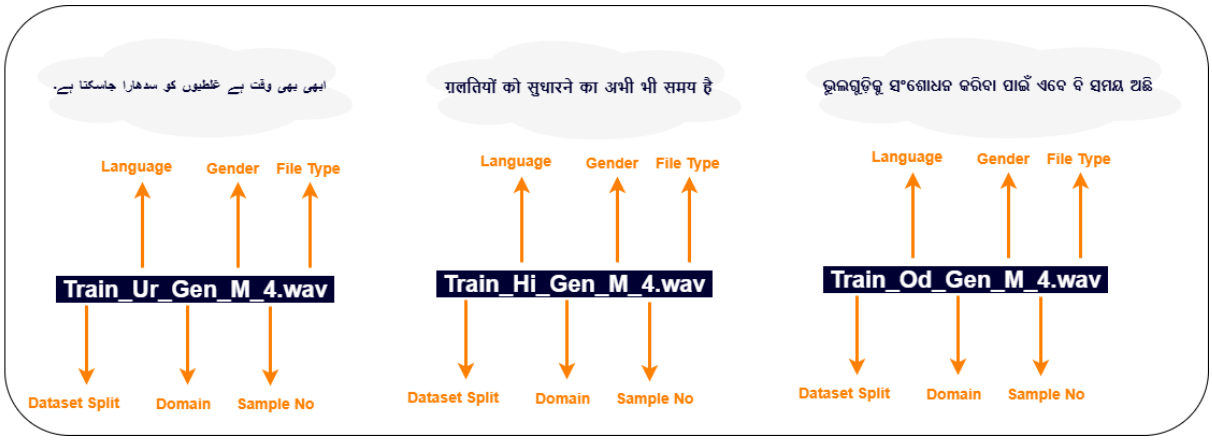

Figure 2: A training sample from the general domain, spoken in a male voice, for all three languages.

ters, emojis, superscripts, and subscripts. Hyphenated word fragments are handled by eliminating hyphens and merging the fragments into a single continuous word.

Abbreviations are expanded into their corresponding full forms (e.g., Mr to mister and Dec to December), and numerical digits are converted into their word equivalents. We further filter the dataset by removing sentence pairs that contain English words in Hindi sentences or Hindi words in Urdu sentences. In addition, parallel pairs consisting of fewer than three words or excessively long sentences are discarded to ensure linguistic consistency and suitability for speech synthesis and translation tasks.

4.3 Speech Synthesis

Speech synthesis is performed using Indic ParlerTTS (Sankar et al., 2025) to generate trilingual Hindi-Urdu-Odia parallel audio data. A fixed prompt "Male speaker, normal speaking rate, high-quality recording with no background noise" is used to ensure controlled and consistent voice generation, and a synthetic male speaker profile is employed to simulate speaker characteristics. For each utterance, the corresponding text serves as input, accompanied by a prompt description specifying gender, speaking rate, and recording quality.

We used the Indic ParlerTTS model to transform tokenized text and prompts into raw audio waveforms at a native 44 kHz sampling rate. To maintain consistency for downstream tasks, these waveforms were converted to NumPy arrays and resampled to a 16 kHz mono WAV format via the Librosa library, specifically utilizing the 'Kaiser-best' algorithm for high-fidelity output. We then organized the audio into a structured parallel corpus where each Urdu, Hindi, and Odia utterance is linked by a common index. This systematic approach to naming and storage, illustrated in Figure 2, ensures the dataset remains clean and easy to navigate.

5 Dataset Statistics

As shown in Table 3, our UrHiOdSynth dataset consists of nearly 10,735 speech-text aligned samples per language. These samples are drawn from three distinct domains, namely Government, General, and Healthcare. The table presents an overview of the speech and text data characteristics. For the speech component, it reports the number of audio files, total duration in hours, and basic duration statistics, including minimum, maximum, and average lengths. For the text component it presents total token counts along with maximum and average sentence lengths. Our UrHiOdSynth speech corpus is created as a three-way parallel corpus, which ensures consistency in the number of audio files and sentences across languages.

The corpus contains 32,205 audio recordings, totaling 75.63 hours of speech, and shows an almost even distribution across languages and domains, with each language contributing approximately 25 hours of data. The audio segments range up to 30.20 seconds in duration, while the average utterance lasts 8.53 seconds. The corpus also includes very short segments, with a minimum duration of 0.09 seconds, which reflect expressions commonly found in conversational and administrative speech.

Sentence-level token analysis shows that the languages in the corpus exhibit moderate syntactic complexity. The dataset contains 5,71,818 tokens in total, with utterances averaging about 17 tokens and reaching a maximum length of 145 to-

Language	Domain	Audio Files	Hours	Audio Duration (s)			Tokens	Token Length	
				Max	Min	Avg		Max	Avg
Hindi	Gov	3,567	7.65	29.20	0.22	7.73	64,080	90	17
	Gen	3,574	7.58	28.52	0.22	7.63	65,346	126	18
	HLT	3,594	8.04	28.20	0.23	8.05	70,894	92	19
Urdu	Gov	3,567	9.35	30.20	0.21	9.43	75,847	126	21
	Gen	3,574	9.04	29.30	0.09	9.10	71,793	145	20
	HLT	3,594	9.27	28.24	0.26	9.29	74,169	101	20
Odia	Gov	3,567	8.16	28.20	0.12	8.24	48,789	79	13
	Gen	3,574	8.17	29.45	0.52	8.23	49,251	106	13
	HLT	3,594	8.36	27.40	0.23	8.38	51,649	77	14
Total	–	**32,205**	**75.63**	**30.20**	**0.09**	**8.53**	**571,818**	**145**	**17**

Table 3: Statistics of our UrHiOdSynth dataset grouped by language and domain. Gen denotes General, Gov denotes Government, and HLT denotes Healthcare. Audio durations are reported in seconds.

Language	General			Government			Healthcare		
	Train (hrs / files)	Test (hrs / files)	Dev (hrs / files)	Train (hrs / files)	Test (hrs / files)	Dev (hrs / files)	Train (hrs / files)	Test (hrs / files)	Dev (hrs / files)
Hindi	4.40 / 2574	1.60 / 500	1.58 / 500	4.82 / 2567	1.50 / 500	1.34 / 500	5.17 / 2594	1.45 / 500	1.42 / 500
Urdu	5.17 / 2574	1.94 / 500	1.92 / 500	5.83 / 2567	1.86 / 500	1.66 / 500	5.96 / 2594	1.66 / 500	1.65 / 500
Odia	4.81 / 2574	1.69 / 500	1.67 / 500	5.07 / 2567	1.66 / 500	1.42 / 500	5.39 / 2594	1.48 / 500	1.49 / 500

Table 4: Domain-wise breakdown of the UrHiOdSynth corpus for the train, development, and test splits, reported by total speech hours and utterance counts per language and domain

kens. Urdu shows slightly longer average utterance lengths, approximately 20 to 21 tokens, than Hindi and Odia, indicating richer morpho-syntactic patterns. The healthcare (HLT) subset consistently exhibits longer utterances than other domains, reflecting the nature of medical language.

For every language, the corpus provides around 3,560–3,600 audio files in each domain, resulting in a balanced distribution across the General, Government, and Healthcare subsets. This balanced setup supports controlled studies of domain adaptation and cross-domain generalization, which are typically difficult to conduct with existing Indic speech datasets such as FLEURS (Conneau et al., 2023).

5.1 Domain-wise data analysis

We follow a domain-consistent data split as presented in Table 4, fixing the development and test sets to 500 utterances per domain and language. This setup allows fair and directly comparable evaluation across all experiments. The amount of training data per language varies between approximately 4.4 and 5.96 hours for each domain, which is adequate for training and benchmarking low-resource speech-to-speech translation (S2ST) systems. While the development and test sets are relatively small, they are constructed to maintain sufficient acoustic and linguistic variability. The Government and Healthcare domains are assigned marginally more training data, reflecting their higher relevance for real-world public-service and clinical translation applications.

5.2 UrHiOdSynth Vs Fleurs

Using the statistics in Table 5, we observe substantial structural and practical differences between FLEURS and UrHiOdSynth. While FLEURS is commonly presented as an n-way parallel multilingual corpus, the actual sentence counts and speech hours vary significantly across languages, with notably limited coverage for low-resource Indic languages such as Hindi, Urdu, and Odia. For instance, FLEURS provides fewer than 5 hours of speech per language and under 2k sentences, which constrains its suitability for training robust end-to-end or direct speech-to-speech translation models. Moreover, FLEURS contains observable noise and lacks systematic human validation, leading to instability during training and challenges in reproducibility.

In contrast, UrHiOdSynth is explicitly designed as a genuinely three-way parallel speech corpus, offering over 23–27 hours of speech and a consistent 10,735 aligned sentences across Hindi, Urdu, and Odia. Crucially, the dataset incorporates human verification to ensure alignment and audio qual-

Languages	UrHiOdSynth		FLEURS	
	Hours	Sentences	Hours	Sentences
Hindi	23.27	10,735	4.82	1,702
Urdu	27.66	10,735	4.64	1,588
Odia	24.70	10,735	3.30	1,327

Table 5: Comparison of dataset scale between UrHiOdSynth and FLEURS in terms of speech duration and number of utterances.

ity, making it substantially more reliable for supervised S2ST training. Another key distinction lies in accent and domain alignment: FLEURS pairs American-accent English with Pakistani-accent Urdu, whereas UrHiOdSynth provides Indic-accent English and Indic-accent Urdu, reflecting realistic deployment conditions for Indic speech technologies. As a result, UrHiOdSynth is not merely larger in scale for these languages, but fundamentally better aligned with the linguistic, acoustic, and sociocultural requirements of Indic speech-to-speech translation systems.

6 Data Validation

After text and speech preprocessing, we use human evaluators for every language pair to guarantee the authenticity of the UrHiOdSynth dataset. Twenty male and female undergraduate and graduate students between the ages of 18 and 30 took part in the validation procedure. We designate two specialists who are fluent in Hindi and either Urdu or Odia and who are skilled in both script and speech for each language pair. Twenty annotators assessed each speech-text pair using a 5-point rating system as part of our human validation procedure. The following is a definition of the scoring criteria: 0 represents total noise, 1 represents unclear voice or text, 2 represents minor noise in the data, 3 represents acceptable quality with few errors, 4 represents data with few misalignments, and 5 represents excellent quality data. Because scores of 3 or higher consistently reflected audible speech and well-aligned text quality, whereas lower values indicated significant mistakes in the data, we only kept voice-text combinations with a score of ≥ 3. Sentences and audio snippets with alignment scores of 1 or 2 are eliminated. We also verified the results with the human assessors used for the particular language combination. The entire dataset text and speech are manually verified. Before a final decision is reached about retention or removal, instances with a score of three are subjected to additional review. This procedure preserves the three-way parallel structure of the dataset while guaranteeing consistency across all language pairs.

7 Experiments and Results

7.1 Model

SeamlessM4T (Barrault et al., 2023) translates spoken utterances across more than one hundred languages, including Hindi, Urdu, and Odia. The model leverages a large-scale corpus of approximately one million hours of open-source speech data and employs the self-supervised w2v-BERT 2.0 framework to learn and capture complex linguistic patterns effectively. The speech-to-text component builds on the Wav2Vec-BERT 2.0 model (Chung et al., 2021), which incorporates additional architectural elements to better capture and model the acoustic cues of spoken words. The system converts speech into text using a modality adapter (Zhao et al., 2022) that bridges speech and textual representations. The translated text then passes to a text-to-unit (T2U) module, which maps it into a sequence of discrete phonetic units. Finally, a HiFi-GAN vocoder (Kong et al., 2020) converts these units into natural-sounding speech.

In our experiments, we fine-tune SeamlessM4T-Medium on both the proposed UrHiOdSynth corpus and the existing FLEURS dataset using an identical training configuration to ensure fair comparison. For each dataset, 80% of the data is used for training and 10% for validation, with the remaining 10% for testing. The model is optimized using the AdamW optimizer with a learning rate of $1e^{-5}$, and gradient clipping with a maximum norm of 0.1 is applied to improve training stability. Model evaluation is performed every 100 training steps, enabling consistent monitoring of convergence and performance across all language pairs and datasets.

7.2 Evaluation Metrics

To evaluate translation quality, we adopt a set of automatic metrics that capture the performance of the model, such as BLEU, ChrF, COMET, and WER. The Bilingual Evaluation Understudy (BLEU) assesses n-gram overlap between generated outputs and reference translations, serving as an indicator of lexical correspondence. Character n-gram F-score (ChrF) computes character-level precision and recall and is particularly well suited to morphologically rich and script-diverse Indic languages. Crosslingual Optimized Metric for Evaluation of

Language Pair	UrHiOdSynth				FLEURS			
	BLEU	ChrF	COMET	WER	BLEU	ChrF	COMET	WER
Hindi–Odia	**17.34**	**45.19**	**0.7523**	**0.7176**	16.58	40.26	0.7161	0.7300
Hindi–Urdu	**17.10**	**43.62**	**0.7498**	**0.6720**	15.70	43.23	0.7037	0.7132
Odia–Urdu	**16.51**	**44.08**	0.6891	0.7340	15.02	41.11	**0.7187**	**0.6947**

Table 6: Performance comparison of SeamlessM4T-Medium on different language pairs using our corpus and the FLEURS corpus.

Translation (COMET), a learned reference-based metric, is employed to better reflect semantic adequacy and fluency, as it has been shown to correlate more closely with human judgments. Additionally, Word Error Rate (WER) is reported to assess the accuracy of generated speech transcriptions, providing insight into pronunciation quality and the robustness of the speech generation pipeline. Collectively, these metrics facilitate a balanced evaluation of both textual and speech-level translation performance.

7.3 Results and Discussion

Table 6 presents a comparative evaluation of SeamlessM4T across three low-resource Indic language pairs using our UrHiOdSynth corpus and the FLEURS benchmark. Overall, the results clearly demonstrate the effectiveness of the proposed corpus in improving speech translation quality, particularly for low-resource Indic languages.

Across Hindi to Odia and Hindi to Urdu, UrHiOdSynth outperforms FLEURS on all major MT metrics. In particular, Hindi–Odia shows gains of +0.76 BLEU, +4.93 ChrF, and +0.036 COMET, along with a lower WER, indicating not only improved lexical overlap, but also better semantic adequacy and robustness to speech recognition. A similar pattern is observed for Hindi to Urdu, with the proposed corpus achieving higher BLEU, ChrF and COMET scores along with a lower WER.

For Odia to Urdu, the results reveal a more nuanced behavior. While UrHiOdSynth achieves higher +1.49 BLEU and +2.97 ChrF scores, FLEURS slightly outperforms in COMET and WER. This divergence highlights that lexical similarity metrics alone are insufficient to capture semantic fidelity and recognition quality, especially for linguistically distant pairs lacking direct phonetic or script-level overlap. The stronger COMET and WER scores on FLEURS suggest that broader multilingual exposure can still benefit semantic consistency in challenging low-resource language translations.

7.4 Domain-wise Performance Analysis

Table 7 reports the domain-wise performance of SeamlessM4T on the UrHiOdSynth dataset across general, government, and healthcare domains. Hindi–Odia achieves the strongest results in the general domain, with the highest BLEU (19.56) and ChrF (46.14). This means the model handles common, wide-ranging vocabulary effectively. The COMET achieves its highest value in the government domain (0.7671) despite lower BLEU and ChrF. It shows that the model is able to preserve the semantics of the utterances despite reduced lexical and character-level overlap with the references. The lower WER (0.7281) further suggests that speech generation is accurate for government domain data, which is generally structured and formally expressed.

For Hindi–Urdu, the healthcare domain clearly dominates, yielding the best scores across all metrics (BLEU 20.68, ChrF 53.11, COMET 0.7416, WER 0.5527). This consistent improvement suggests that healthcare speech exhibits higher terminological consistency and shared lexical roots between Hindi and Urdu, benefiting both semantic modeling and acoustic generation. The substantially lower WER highlights that the model produces cleaner and more intelligible speech in this domain, a critical requirement for safety-sensitive applications.

In contrast, Urdu–Odia shows relatively stable but lower performance across domains, with its best results concentrated in the general domain. The drop in BLEU and ChrF for government and

Language Pair	GENERAL				GOVERNMENT				HEALTHCARE			
	BLEU	ChrF	COMET	WER	BLEU	ChrF	COMET	WER	BLEU	ChrF	COMET	WER
Hindi–Odia	**19.56**	**46.14**	0.7421	0.7631	16.48	44.82	**0.7671**	**0.7281**	17.21	45.19	0.7232	0.7456
Hindi–Urdu	18.54	45.54	0.7096	0.6802	16.00	43.67	0.6728	0.7119	**20.68**	**53.11**	**0.7416**	**0.5527**
Urdu–Odia	**17.16**	**45.17**	**0.7281**	0.7123	16.82	44.21	0.7202	**0.6874**	16.22	44.02	0.7133	0.6941

Table 7: Domain-wise performance of SeamlessM4T across different language pairs on UrHiOdSynth dataset.

healthcare domains indicates increased difficulty in mapping domain-specific terminology between two linguistically and script-divergent languages. However, the comparatively stable COMET scores suggest that the model still captures coarse semantic intent even when lexical fidelity degrades.

Overall, these results strongly indicate that domain-aware training and evaluation are indispensable for S2ST systems in Indian language contexts. The observed domain-specific gains, particularly in healthcare, suggest that targeted corpus design and domain-adaptive fine-tuning could yield substantial improvements. From a deployment perspective, this is not optional: models intended for governance or healthcare must be evaluated within those domains to ensure reliability, intelligibility, and semantic correctness.

In summary, Table 7 demonstrates that SeamlessM4T exhibits meaningful domain-dependent behavior, with performance governed by both linguistic proximity and domain regularity. This finding provides strong empirical motivation for future work on domain-adaptive S2ST models and evaluation protocols tailored to low-resource, multilingual settings.

8 Conclusion and Future works

We present our UrHiOdSynth, a three-way parallel speech corpus for Hindi, Urdu, and Odia with balanced coverage across General, Government, and Healthcare domains. The dataset enables controlled evaluation of domain effects in speech-to-speech translation and related speech–text tasks. Under identical experimental settings with SeamlessM4T, evaluation on UrHiOdSynth yields higher BLEU, ChrF, and COMET scores for most language pairs compared to FLEURS, while achieving comparable or lower WER, demonstrating the dataset's suitability for systematic S2ST analysis.

Domain-wise results show clear performance variation across application domains, with structured domains such as government and healthcare exhibiting distinct metric trends across language pairs. These findings emphasize the importance of domain-aware evaluation for low-resource speech translation. Future work will extend UrHiOdSynth to additional Indic languages, including Maithili, Bengali, Kashmiri, Bhojpuri, and others. We also plan to explore domain-adaptive training strategies and direct speech-to-speech translation models, and to incorporate human evaluation to better assess real-world usability.

9 Limitation

Despite representing a significant advancement in synthesizing parallel speech data for direct speech translation, our approach has limitations. The synthesis process depends on existing parallel text data, making the quality of the synthesized speech sensitive to the original data. In addition, synthetic speech often lacks the natural variability of human speech, such as variations in tone, rhythm, and emotion. Furthermore, synthetic speech disregards background noise, fillers, and pauses, all of which are common elements of natural speech. Since synthetic data is frequently generated using a small number of voices, it fails to capture the diversity of real speakers with different accents, speaking styles, and recording environments.

10 Ethics Statement

This study adheres to ethical research practices in the creation and use of multilingual speech data. The textual data consists of cleaned and validated parallel sentences and does not contain any personally identifiable or sensitive information. All speech data are synthetically generated using IndicTTS, avoiding concerns related to speaker consent, privacy, or voice misuse. For Odia text generation, Hindi text is translated using the Google Translate API strictly for research purposes. The resulting Urdu–Hindi–Odia parallel speech corpus is intended solely for academic and non-commercial use.

References

Ebrahim Ansari, MH Sadreddini, Mostafa Sheikhalishahi, Richard Wallace, and Fatemeh Alimardani. 2017. Using english as pivot to extract persian-italian parallel sentences from non-parallel corpora. *arXiv preprint arXiv:1701.08339*.

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*.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2019. Librivoxdeen: A corpus for german-to-english speech translation and german speech recognition. *arXiv preprint arXiv:1910.07924*.

Soham Bhattacharjee, Mukund K Roy, Yathish Poojary, Bhargav Dave, Mihir Raj, Vandan Mujadia, Baban Gain, Pruthwik Mishra, Arafat Ahsan, Parameswari Krishnamurthy, and 1 others. 2025. Coril: Towards enriching indian language to indian language parallel corpora and machine translation systems. *arXiv preprint arXiv:2509.19941*.

Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kaushal Santosh Bhogale, Deovrat Mehendale, Tahir Javed, Devbrat Anuragi, Sakshi Joshi, Sai Sundaresan, Aparna Ananthanarayanan, Sharmistha Dey, Anusha Srinivasan, Abhigyan Raman, and 1 others. 2025. Towards bringing parity in pretraining datasets for low-resource indian languages. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

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.

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging small multilingual corpora for smt using many pivot languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1192–1202.

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*.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34:15748–15761.

Qingkai Fang, Yan Zhou, and Yang Feng. 2023. Daspeech: Directed acyclic transformer for fast and high-quality speech-to-speech translation. *Advances in Neural Information Processing Systems*, 36:72604–72623.

Zairan Gong, Xiaona Xu, and Yue Zhao. 2025. Tibetan–chinese speech-to-speech translation based on discrete units. *Scientific Reports*, 15(1):2592.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2025a. Benchmarking hindi-to-english direct speech-to-speech translation with synthetic data. *Language Resources and Evaluation*, pages 1–39.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2025b. Direct speech-to-speech neural machine translation: A survey. *Speech Communication*, page 103317.

Zejiang Hou, Daniel Garcia-Romero, and Kyu J Han. 2025. Hyper-adapter for parameter-efficient multilingual asr adaptation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kenji Imamura, Masao Utiyama, and Eiichiro Sumita. 2023. Pivot translation for zero-resource language pairs based on a multilingual pretrained model. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 348–359.

Tahir Javed, Kaushal Bhogale, and Mitesh M Khapra. 2025. Nirantar: Continual learning with new languages and domains on real-world speech data. *arXiv preprint arXiv:2507.00534*.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Towards building asr systems for the next billion users. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, pages 10813–10821.

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M Khapra. 2023. Svarah: Evaluating english asr systems on indian accents. *arXiv preprint arXiv:2305.15760*.

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

Robert Jones, Firstname2 Lastname2, and Firstname3 Lastname3. 2022. An excellent paper introducing the ABC toolkit. In (Smith et al., 2022), pages 105–109.

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.

Sireesh Haang Limbu. 2020. Direct speech to speech translation using machine learning.

Anindita Mondal, Anil Kumar Vuppala, and Chiranjeevi Yarra. 2024. Iiit-speech twins 1.0: An english-hindi parallel speech corpora for speech-to-speech machine translation and automatic dubbing. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayuth Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2024. Translatotron 3: Speech to speech translation with monolingual data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10686–10690. IEEE.

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, Yoach Lacombe, Sherry Thomas, Praveen Srinivasa Varadhan, Sanchit Gandhi, and Mitesh M Khapra. 2025. Rasmalai: Resources for adaptive speech modeling in indian languages with accents and intonations. *arXiv preprint arXiv:2505.18609*.

Mohammad Sarim, Saim Shakeel, Laeeba Javed, Mohammad Nadeem, and 1 others. 2025. Direct speech to speech translation: A review. *arXiv preprint arXiv:2503.04799*.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. 2025. Indicsynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22037–22060.

Jane Smith, Firstname2 Lastname2, and Firstname3 Lastname3. 2022. A really good paper about Dynamic Time Warping. In *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pages 100–104, Incheon, Korea.

Abhimanyu Talwar and Julien Laasri. 2025. Pivot language for low-resource machine translation. *arXiv preprint arXiv:2505.14553*.

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*.

Zhanglin Wu, JiaXin Guo, Daimeng Wei, Zhiqiang Rao, Zongyao Li, Hengchao Shang, Yuanchang Luo, Shaojun Li, and Hao Yang. 2024. Improving the quality of iwlst 2024 cascade offline speech translation and speech-to-speech translation via translation hypothesis ensembling with nmt models and large language models. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 46–52.

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*.