VAANI: A Spontaneous, Multi-Speaker Speech-to-Speech Translation Benchmark for Indian Languages

Anonymous ACL submission

Abstract

Speech-to-Speech (S2S) translation systems frequently exhibit severe performance degradation in real-world environments, particularly within linguistically complex contexts characterized by pervasive code-mixing, such as those found in India. To address this critical gap, we introduce VAANI, a novel parallel speech corpus designed for direct S2S translation across five major Indian languages – Hindi, Bengali, Telugu, Gujarati, and Marathi. Unlike existing corpora that rely on formal read speech, VAANI uniquely captures the nuances of spontaneous communication, encompassing diverse speakers, regional dialects, and real-world domains. We comprehensively benchmark this dataset against leading speech foundation models, including the SeamlessM4T, AudioPaLM, and Whisper families. Whisper-Large-v3 achieves the strongest overall results in our evaluations. Our acoustic and spectral analyses expose a persistent generation gap in current architectures – while natural spontaneous speech exhibits dense, continuous energy distributions, artificially generated speech suffers from distinct spectral sparsity and reduced acoustic richness. VAANI provides a rigorous, ecologically valid testbed for evaluating and advancing robust, multilingual S2S systems for real-world deployment.[1]

1 Introduction and Prior Art

Speech-to-Speech (S2S) translation is the complex task of converting spoken utterances in a source language directly into spoken output in a target language (Kakwani et al., 2020; Dhawan, 2022; Sarim et al., 2025; Yeh, 2014). Early approaches to this challenge relied heavily on cascaded pipelines, sequentially integrating Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) systems (Kumar et al., 2023; Macháček, 2024; Dutoit, 1997; Sasirekha and Chandra, 2012; Kumar et al., 2023; Jain et al., 2024; Hasanabadi, 2023). While foundational, these cascaded models suffer from compounding errors across subsystems, introduce significant latency, and critically fail to preserve paralinguistic information. Elements such as prosody, intonation, rhythm, speaker traits, and emotion play a vital role in meaning preservation and naturalness (Abdullah et al., 2021; Guo et al., 2023; Oyucu, 2023). These limitations necessitate a paradigm shift toward direct S2S translation models (Trancoso et al., 2023; Kim et al., 2024; i Barrobes, 2006; Abdel Monem et al., 2008; Trivedi et al., 2018; Pitrelli et al., 2006). Unlike their cascaded counterparts, direct models map source speech to target speech without relying on intermediate textual representations (Etchegoyhen et al., 2022; Gregori, 2021; Kim et al., 2024; Hasanabadi, 2023; Trivedi et al., 2018). However, training these architectures requires massive, high-quality parallel speech corpora – a resource that remains acutely scarce, particularly for multilingual and low-resource settings like India (Opitz et al., 2025; Matuschek and Gurevych, 2013; Anderson and Fenichel, 1989; Hasanabadi, 2023; Oyucu, 2023; Ifeanyi et al., 2014).

Existing initiatives, such as IndicT-EDST (Sethiya et al., 2024), successfully address the scarcity of Speech-to-Text (S2T) data but fall short of providing the large-scale parallel corpora required for end-to-end S2S translation in Indian languages (Abdullah et al., 2021). Similarly, while foundational models like SeamlessM4T support S2S capabilities, they only accommodate a highly restricted subset of Indian languages (Telugu, Hindi, Urdu, Bengali) as target outputs (Mujadia and Sharma, 2023; Greenberg, 1997; Jain et al., 2024). Furthermore, widely adopted benchmarks like FLEURS pri-

[1] We are unable to include VAANI as supplementary due to the anonymity policy. We are committed to open-source it upon acceptance of the paper.

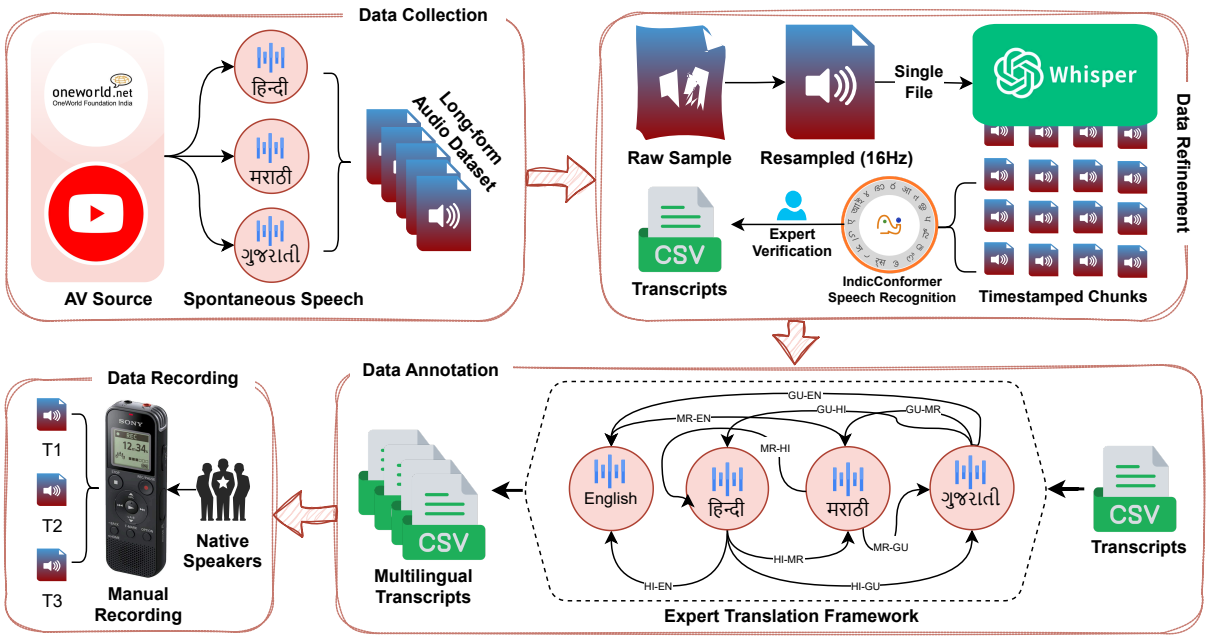


Figure 1: **Overview of the VAANI dataset creation pipeline.** The illustrated methodology (shown here for Hindi, Marathi, and Gujarati) is applied systematically across all five languages. The workflow involves: (1) **Data Collection** of spontaneous, real-world speech; (2) **Data Refinement** via ASR segmentation and expert human validation; (3) **Data Annotation** to translate utterances into all corresponding target languages; and (4) **Data Recording** by native speakers in controlled environments to yield the final speech-to-speech corpus.

marily evaluate models on "read speech." While acoustically clean and highly controlled, read speech fails to reflect the realities of spontaneous human communication (Greenberg, 1997; Jain et al., 2024). Natural discourse is characterized by hesitations, repetitions, repairs, informal expressions, dialectal variations, and frequent code-switching–phenomena that heavily define the Indian multilingual context. Because models trained or evaluated strictly on read speech suffer substantial performance degradation when exposed to real-world spontaneity (Jain et al., 2024), there is a critical need for a spontaneous, dialect-diverse, and parallel S2S benchmark tailored to Indic languages. We summarize the specifications of prominent existing speech translation datasets in Table 1. Extended related work is discussed in Appendix A.

Recent studies attempt to bridge this divide through advanced modeling techniques (Zhang et al., 2024). Fine-tuned models like SeamlessM4T-v2-large have driven significant progress in multilingual speech processing for Indian languages (Gupta, 2022), and systems trained on massive S2T corpora demonstrate improved transcription performance (Omazić and Lekić, 2021). Despite these advancements, a substantial gap persists in both resources and foundational models capable of handling spontaneous, conversational, and code-switched Indic speech for direct S2S translation (Pilz and Schneider, 2025).

To bridge this gap, we introduce VAANI, a spontaneous S2S translation dataset encompassing five major Indian languages – Hindi, Bengali, Telugu, Gujarati, and Marathi. Motivated by their geographical proximity, phonological similarity, and high bilingualism rates (Sherborne, 2024; Codenie et al., 1997; Sutcliffe and Maiden, 2002), these languages provide an ideal testbed for analyzing cross-lingual acoustic alignment while mirroring authentic Indian communication patterns. VAANI comprises approximately 10 hours of parallel speech derived from real-world conversations, interviews, panel discussions, and public talks. We segment this audio into discrete utterances, systematically aligning each with gold-standard source transcripts, text translations, and gold-standard speech translations across all five languages. Unlike existing corpora, VAANI explicitly emphasizes: (i) **Spontaneity:** The natural inclusion of conversational disfluencies, hesitations, and acoustic markers, (ii) **Code-switching:** The preservation of frequent intra- and inter-sentential language mixing, (iii) **Dialectal Diversity:** A broad representation of regional

Table 1: Comparison of VAANI with existing Speech-to-Speech (S2S) and multilingual translation datasets.

Dataset	Specification
S2ST-Dialect (Sood et al., 2025)	Speech-to-Speech machine translation (S2ST) for four Hindi dialects, i.e., Awadhi, Bhojpuri, Braj Bhasha, and Magadhi.
IndicST (Sankar et al., 2025)	The largest speech translation dataset for Indian languages, spanning over 44 thousand hours of audio and 17 million aligned text segments across 14 Indian languages and English.
CVSS (Jia et al., 2022)	A Massively Multilingual-to-English Speech-to-Speech Translation Corpus.
Waxal (Diack et al., 2026)	A new large-scale resource for 27 African languages, integrating both ASR (spontaneous speech) and high-fidelity TTS data to enable full-duplex conversational systems.
VAANI (Our Work)	Speech-to-speech and text-to-text translation datasets for five Indian languages for benchmarking speech-based LLMs.

speaker variations, and (iv) **Conversational Settings:** Multi-speaker, real-world dynamics rather than isolated, single-speaker read prompts.

Using VAANI as a gold-standard reference, we systematically benchmark state-of-the-art speech foundation models, including SeamlessM4T-Large-v2, SeamlessM4T-Large, AudioPaLM-2, Whisper-Large-v2, and Whisper-Large-v3. We evaluate textual fidelity using a comprehensive suite of metrics: ROUGE-2 F1, ROUGE-L F1, METEOR, BLEU, BERTScore, Word Error Rate, Phoneme Error Rate (PER), and Character Error Rate. Among the evaluated systems, Whisper-Large-v3 demonstrates the strongest overall performance, outperforming Whisper-Large-v2 by 2%–4% in BERTScore across the language pairs.

We conduct a rigorous analysis of the acoustic similarity between the source and translated speech. By measuring Mel Cepstral Distortion (MCD), Pitch, Duration, and Energy correlations, alongside Speaker Embedding Cosine Similarity, Mean Opinion Score (MOS), and the Structural Similarity Index (SSIM), we uncover distinct synthesis patterns. We observe that the Hindi-Marathi pair yields the highest cross-lingual duration, energy, and pitch correlations (0.7542, 0.7509, and 0.7577, respectively), strongly reflecting their underlying phonological proximity. Our spectrogram analysis exposes a persistent flaw in current models – while natural speech exhibits dense, continuous energy distributions, artificially generated speech suffers from distinct spectral sparsity, indicating a significant loss of acoustic richness.

2 Dataset

In this section, we detail the methodology for constructing our dataset, as illustrated in Figure 1. Our pipeline comprises four primary stages – data collection, data refinement, data annotation, and recording. To ensure high-fidelity outputs, we rigorously validate the data at the conclusion of each stage before progressing to the next. We outline these key steps in detail below.

2.1 Data Collection

Our goal is to create a diverse, spontaneous speech-to-speech translation dataset for Hindi, Bengali, Telugu, Gujarati, and Marathi. We begin by collecting large-scale raw speech data across these languages as the foundation of the annotated corpus, focusing on three key dimensions:

Spontaneous Speech and Code-Mixing. To ensure our dataset reflects the natural complexities of real-world communication, we prioritize spontaneous speech over traditional, script-based "read speech." Read speech usually relies on monolingual text and lacks natural prosody. In contrast, spontaneous speech naturally captures prevalent linguistic phenomena such as code-mixing, which is a defining characteristic of Indic language usage. To capture these authentic dynamics, we curate a collection of real-life utterances from interviews, panel discussions, and public speeches. We source all audio from Creative Commons-licensed platforms, primarily YouTube and the EDAA portal, an open-access repository for community-driven audio content.[2]

Diverse Speakers. India's linguistic landscape includes rich dialectal variation, often underrepresented in existing speech translation datasets that focus on standard forms. To address this, we collect data across diverse regional dialects and ensure

[2]Ek Duniya Anek Awaz Portal.

Table 2: **Statistics of VAANI in terms of (a) average speech duration and (b) average words per transcript.** Each row corresponds to a base language, and columns denote target languages. The number of utterances per base language are as follows – Hindi (965), Gujarati (1320), Marathi (910), Telugu (1010), and Bengali (908).

(a) Average Speech Duration (seconds)

	Hindi	**Gujarati**	**Marathi**	**Telugu**	**Bengali**
Hindi	14.05	13.56	15.55	15.51	3.59
Gujarati	7.96	7.42	8.25	7.99	8.48
Marathi	16.67	15.54	18.06	6.77	10.53
Telugu	16.67	15.54	11.26	18.06	13.22
Bengali	8.56	9.62	8.49	8.92	8.44

(b) Average Number of Words per Transcript

	Hindi	**Gujarati**	**Marathi**	**Telugu**	**Bengali**
Hindi	12.55	11.25	16.74	12.45	13.44
Gujarati	8.76	9.15	8.12	8.95	8.10
Marathi	16.69	13.44	16.55	13.83	16.69
Telugu	13.46	12.40	11.62	13.32	11.96
Bengali	12.27	12.95	9.56	9.94	15.25

speaker demographics reflect a broad, inclusive range of socioeconomic backgrounds.

Domain Diversity. As we aim to create speech technologies relevant to rural India, we focus heavily on domains with high local significance, such as agriculture and healthcare. We maintain a balanced distribution across these domains to ensure broad topical coverage, while additionally prioritizing data that discusses localized regional issues.

2.2 Data Refinement

We initially collect the raw data as continuous, long-form audio files. Because translating unbroken long-form audio introduces significant complexity, we transform and standardize this data to establish precise temporal alignment between the source audio, its transcription, and the eventual target translation. We execute the following refinement pipeline:

Audio Resampling. We resample all audio files to 16 kHz to guarantee compatibility with modern ASR architectures.

Timestamp Prediction. We segment the long-form audio into discrete, manageable utterances. We utilize the Whisper ASR model specifically to generate accurate timestamp boundaries for these audio chunks.

Transcription and Human Validation. During our initial evaluations, we observe that Whisper's zero-shot transcription quality for these Indic languages is suboptimal. We deploy IndicConformer [3], a model explicitly optimized for Indian language ASR, to transcribe the Whisper-segmented audio chunks. To ensure absolute fidelity to the source audio, we subject all machine-generated transcripts to rigorous human validation. Our in-house language experts review the ASR outputs against the original audio using Audacity. They manually adjust timestamp boundaries and correct the text to produce our final gold-standard transcripts.

During this refinement, we execute several critical modifications: (i) Correcting ASR spelling and word recognition errors, (ii) Inserting appropriate punctuation to reflect conversational flow, (iii) Standardizing conversational disfluencies, such as false starts, self-corrections, and filler words, (iv) Formatting numbers and standardizing abbreviations, and (v) Accurately annotating and formatting instances of code-mixing.

2.3 Data Annotation

Following the refinement and validation of the source audio chunks and their gold-standard transcripts, we initiate the translation phase to construct our multi-way parallel corpus. We systematically translate the validated text across our source languages – Bengali, Hindi, Gujarati, Marathi and Telegu into the remaining target languages. This exhaustive cross-lingual mapping ensures that every source utterance aligns with high-quality, human-validated text translations in multiple target languages, forming the backbone of our dataset.

2.4 Audio Recording

To build the S2S dataset, we record target-language audio for translated segments using native speakers in a structured reading task. Sessions are conducted in acoustically controlled environments, with recordings captured on a Sony ICD-PX470 at 16 kHz for consistency.

2.5 Dataset Statistics

We present descriptive statistics of VAANI in Table 2, summarizing the distributional properties of speech data across base languages and their translated counterparts. For each base language, we report the total number of utterances, the mean speech duration (in seconds), and the mean transcript length measured as the number of words.

[3] https://github.com/AI4Bharat/IndicConformerASR

Statistics are provided for the base language as well as for the corresponding translated versions, enabling a comparative analysis of temporal and lexical variations introduced through translation. Marathi exhibits the longest average utterance durations and highest word counts, reflecting its relatively verbose spoken form, while Gujarati shows consistently shorter utterances and lower lexical counts across both base and translated conditions. Hindi occupies an intermediate position in both duration and word length. We observe that translation does not uniformly preserve these characteristics. For several language pairs, translated speech differs measurably from the base language in both duration and lexical density. These variations highlight the need to account for language-specific and translation-specific effects when modeling multilingual and cross-lingual speech data.

To facilitate demographic comparisons, we assigned sentence-level gender labels using the Common-Voice-Gender-Detection model (Sakthi, 2025). As detailed in Table 3, gender distributions vary by language, with Marathi skewing male and Gujarati skewing female. To evaluate linguistic complexity, we measured Mean Sentence Length (MSL), Mean Segmental Type-Token Ratio (MSTTR), Mean Morphological Complexity (MMC), Mean Clause Depth (MCD), and Average Weighted Sentence Complexity (AWSC). A Spearman rank correlation analysis (Table 4) confirms these metrics capture distinct linguistic dimensions. MMC operates independently of syntactic length ($r = 0.027, p < 0.05$), while MSTTR notably correlates negatively with MSL ($r = -0.669$), reflecting decreased lexical diversity in longer segments. Conversely, MSL and AWSC are nearly perfectly correlated ($r = 0.993, p < 0.001$), suggesting that syntactic length heavily overlaps with the composite AWSC score in these languages. Kruskal-Wallis tests (Table 5) demonstrate that language and topical domain are the primary drivers of linguistic variation, significantly impacting all metrics ($p < 0.001$). Visualizing these distributions (Figures 5–7) reveals that technical domains (e.g., Healthcare, Government Policy) predictably constrain lexical diversity (MSTTR), while certain languages, like Marathi, maintain a higher baseline for syntactic complexity (MCD) across all topics. Conversely, Mann-Whitney U tests indicate that speaker gender has a negligible impact on structural complexity. No statistically significant gender differences were observed in Hindi, Bengali, or Telugu ($p > 0.05$). While female speakers in Gujarati and Marathi exhibited slightly lower MSL and MCD scores than males, the effect sizes were minimal (Cliff's $d \approx 0.15$). Thus, while statistically detectable in specific subsets, gender differences do not practically impact the structural complexity of the corpus compared to the dominant factors of language and domain.

Table 3: **Speaker distribution across languages.** Values denote the number of speakers, with percentages indicating the proportion within each language.

Language	Male	Female	Male (%)	Female (%)
Bengali	774	133	85.3	14.7
Gujarati	773	517	59.9	40.1
Hindi	1144	363	75.9	24.1
Marathi	984	486	66.9	33.1
Telugu	712	258	73.4	26.6
Total	**4387**	**1757**	**71.4**	**28.6**

Table 4: **Spearman correlation coefficients between linguistic metrics.** All reported correlations are statistically significant ($p < 0.05$).

	MSL	MSTTR	AWSC	MMC	MCD
MSL	1.00				
MSTTR	-0.67	1.00			
AWSC	0.99	-0.66	1.00		
MMC	0.03	0.14	0.09	1.00	
MCD	0.64	-0.48	0.70	0.09	1.00

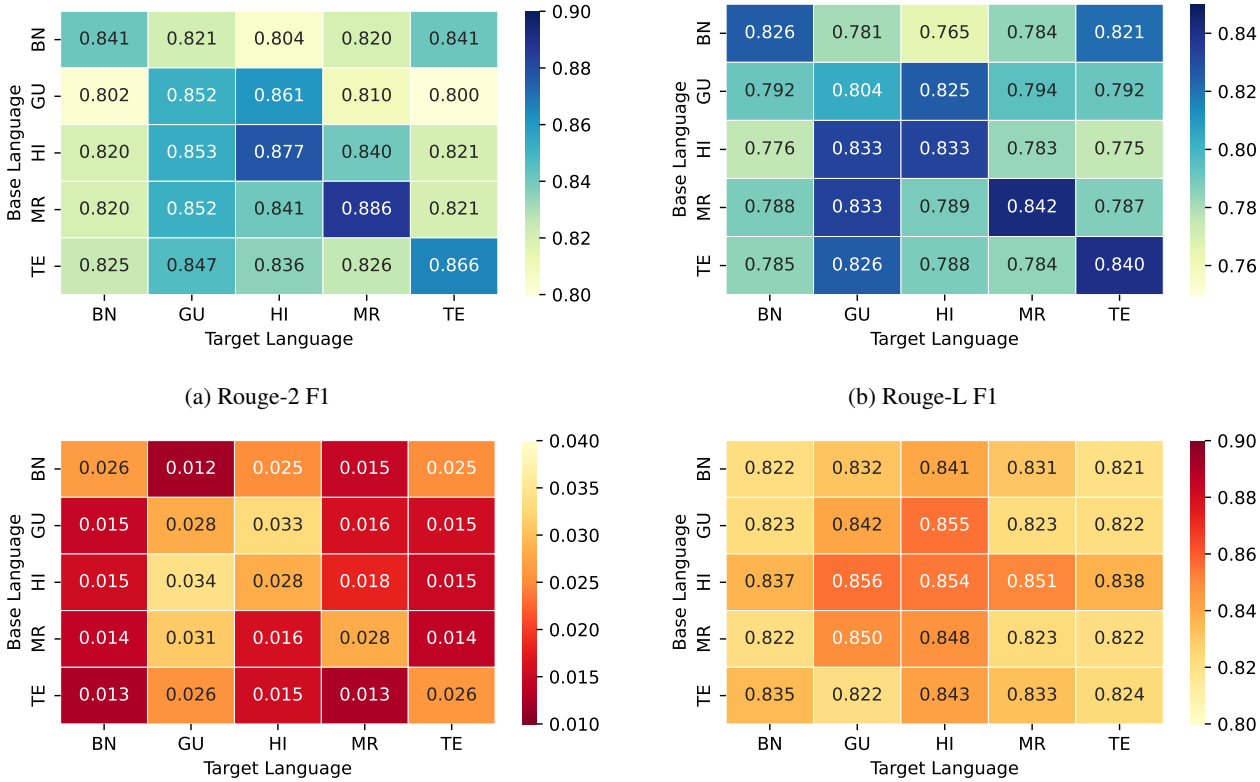
Table 5: **Omnibus Kruskal-Wallis Test**. Kruskal–Wallis analysis revealing pronounced and statistically significant language- and domain-level differences across all linguistic metrics, with uniformly low p-values (< 0.001).

	Language Effect		**Domain Effect**	
Metric	H**-stat**	p**-value**	H**-stat**	p**-value**
MSL	1088.8825	< 0.001	475.2272	< 0.001
MSTTR	709.7583	< 0.001	185.1352	< 0.001
AWSC	904.6287	< 0.001	475.7761	< 0.001
MMC	1845.7636	< 0.001	590.0622	< 0.001
MCD	571.9831	< 0.001	287.4801	< 0.001

3 Methodology

3.1 Multilingual Corpus Construction

To systematically evaluate multilingual speech-text alignment and cross-lingual transfer, we constructed a multi-way parallel dataset encompassing five Indic languages – Gujarati (GU), Hindi (HI), Marathi (MR), Bengali (BN), and Telugu (TE).



(a) Rouge-2 F1

(b) Rouge-L F1

(c) Word Error Rate

(d) BERTScore

Figure 2: **Cross-lingual performance across five Indian languages – Bengali (BN), Gujarati (GU), Hindi (HI), Marathi (MR), Telugu (TE).** Rows are base languages, columns are target languages. Cross-lingual transfer drops modestly, especially between less related languages. BERTScore stays stable (semantic consistency), and lower WER indicates fewer token errors.

For every source utterance, comprising the original speech signal and its native transcript, we generate parallel text transcripts in the target languages via expert translation. For example, a native Gujarati speech sample and its transcript are aligned with newly translated Hindi and Marathi texts, while strictly preserving the original Gujarati acoustic signal. This permutation is applied symmetrically across the source languages, resulting in a densely aligned, one-to-many multimodal corpus where every speech sample is grounded by text references in multiple languages.

3.2 Benchmarking State-of-the-Art Models

We benchmark our gold-standard, human-annotated transcripts against outputs generated by a suite of state-of-the-art speech foundation models. The evaluated models include SeamlessM4T-large-v2, SeamlessM4T-large-v1, AudioPaLM-2, Whisper-large-v2, and Whisper-large-v3. Our evaluation protocol was designed to rigorously test both the monolingual transcription (ASR) and cross-lingual translation (S2T/S2ST) capabilities of these architectures.

3.3 Evaluation

Model performance is quantified using the following metrics:

Textual Translation Quality. To measure the lexical and semantic fidelity of the generated text against human references, we computed a robust set of automated metrics: ROUGE-2 F1 (R2-F1), ROUGE-L F1 (RL-F1), METEOR, BLEU, and BERTScore (Appendix C).

Acoustic and Prosodic Fidelity. To assess cross-lingual speech-to-speech translation quality, we compared the original speech signals with their synthesized counterparts. This acoustic evaluation utilized Mel Cepstral Distortion (MCD), prosodic correlations (pitch, duration, and energy), speaker-embedding cosine similarity, Structural Similarity Index (SSIM), and human-evaluated Mean Opinion Scores (MOS).

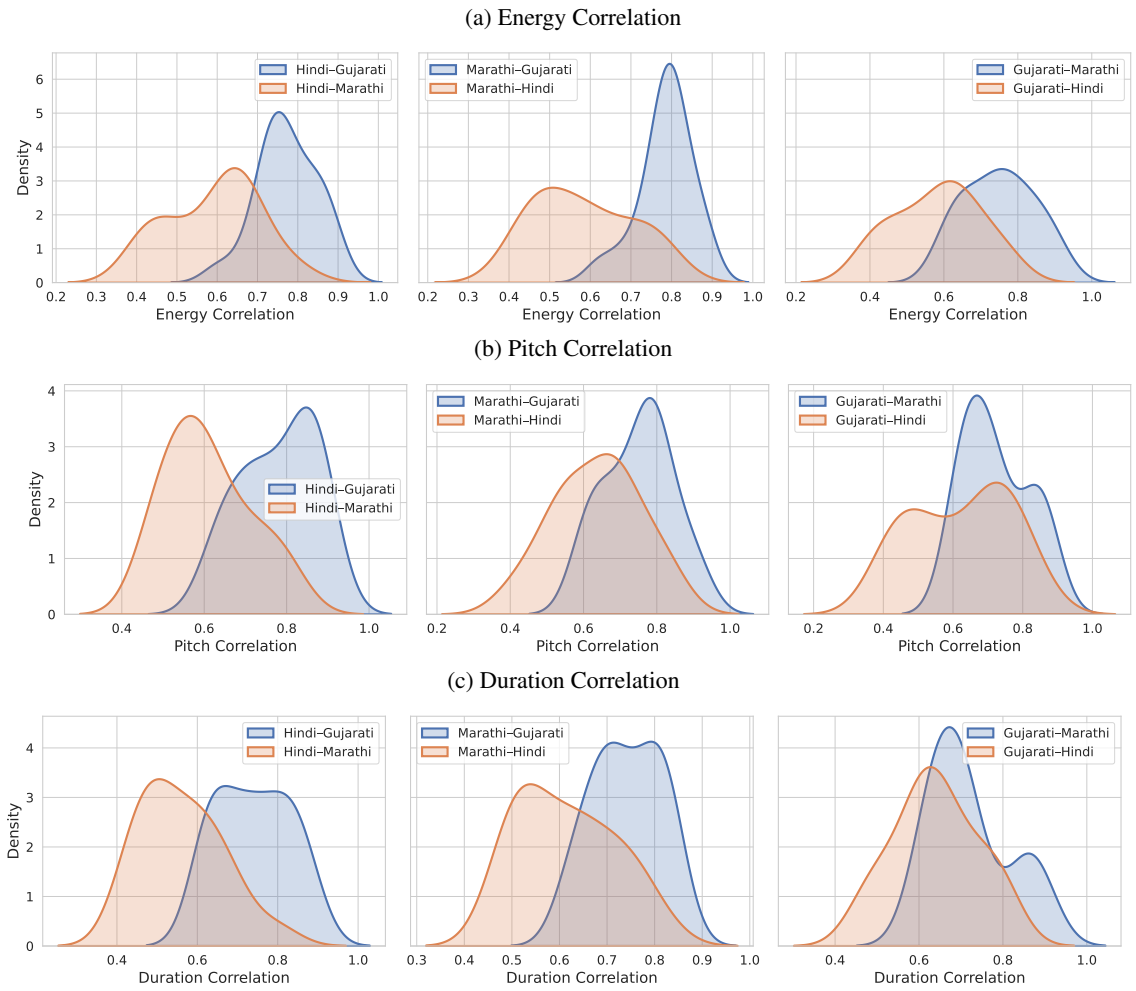


Figure 3: **Distribution of correlation coefficients across selected language pairs.** Kernel density estimates of correlation distributions for selected language pairs.

4 Results and Analysis

Following the evaluation protocol established in Section 3, we report the performance of the benchmarked foundation models on `VAANI`.

Textual and Translation Efficacy. Tables 8, 9, 10, 11 and 12 detail the transcription and translation performance across all source-target language pairs for different LLMs. *Whisper-large-v3* consistently outperforms the other baselines, emerging as the most robust architecture for speech-to-text generation on this dataset. Table 7 reports the performance for cross-lingual translation across different language pairs. Results indicate that translations into Hindi consistently achieve the strongest performance across most evaluation metrics. Translations into Gujarati and Marathi also demonstrate competitive scores, reflecting strong fluency and structural preservation, with minor variations in phrasing accounting for small reductions in alignment. Translations to other Indian languages show slightly higher error rates, indicating good but not perfect recognition accuracy within the same language. This means that while the model maintains a consistent performance trend across different languages, there are still noticeable variations depending on the specific language pair involved.

Acoustic and Prosodic Fidelity. Beyond textual accuracy, we analyze the prosodic alignment between the ground-truth and synthesized audio. Figure 3 presents the distributions of energy, pitch, and duration correlations. A notable finding is the pronounced energy correlation observed in the Gujarati-Marathi translation pair compared to the Gujarati-Hindi pair, suggesting that current models capture prosodic nuances more effectively between certain linguistically proximate languages. Spectral analysis corroborates these prosodic trends. When comparing Hindi to the other two, it con-

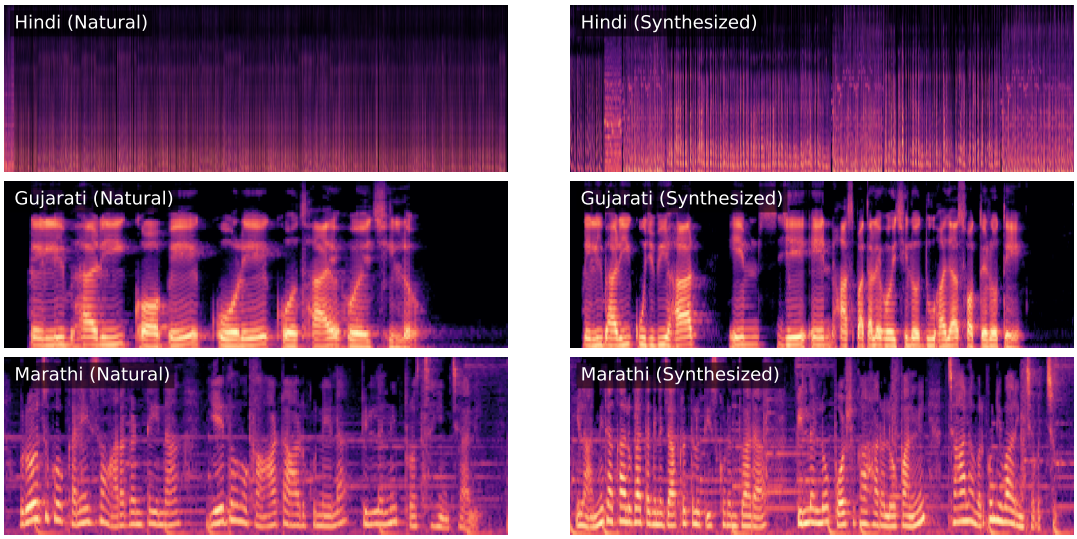


Figure 4: **Spectrogram comparison of natural and synthesized speech for Hindi, Gujarati and Marathi.** Natural speech (left) and generated speech (right) are visualized using log-mel spectrograms.

sistently demonstrates a higher correlation with Gujarati than with Marathi. While all are related Indo-Aryan languages, the acoustic evidence identifies Gujarati and Marathi as the most prosodically similar pair, reflecting shared intonation, energy, and durational patterns that distinguish them from Hindi. Figure 4 side-by-side phonetic visualization for Hindi, Gujarati and Marathi, illustrating how well the models preserve underlying formant structures and spectral integrity. The difference in the visual texture of spectrograms, where artificially generated speech shows more dark spots and natural speech appears brighter and more filled-in, stems directly from the fundamental processes of sound production (Appendix B).

Speech-based Metrics. Table 6 details the speech-based objective metrics for cross-lingual transfer across the five evaluated languages. MCD scores remain relatively stable across most transfers, generally clustering between 2.77 and 3.06. This indicates an adequate preservation of the spectral envelope across different base-target pairs, with the notable exception of the Hindi-to-Bengali transfer, which exhibits a higher MCD of 3.56, suggesting greater spectral degradation. When examining prosodic fidelity (Pitch, Duration, and Energy Correlation) alongside Speaker Embedding Cosine Similarity, a distinct bifurcated pattern emerges across the evaluations. For any given base language, the target languages typically split into a high-performing tier (∼0.74–0.77) and a moderate-performing tier (∼0.60-0.66). For example, a Hindi base model transfers its prosodic and speaker characteristics highly effectively to Marathi and Bengali (0.75), but shows weaker correlation when transferring to Gujarati and Telugu (0.65). SSIM scores reflect moderate spectrogram reconstruction fidelity – typically ranging from 0.40–0.50 for most language pairs – underscoring that while zero-shot cross-lingual acoustic transfer is viable, its efficacy is heavily dependent on the specific base-target language pairing.

5 Conclusion and Future Work

We introduced VAANI, the first spontaneous, multi-speaker S2S translation benchmark for five major Indian languages. By capturing real-world phenomena – code-switching, dialects, and disfluencies, it offers a rigorous testbed for modern speech models. Our benchmarking establishes a clear performance hierarchy, with Whisper-Large-v3 leading speech-to-text, outperforming Whisper-Large-v2 by 2%–4% in BERTScore while maintaining lower error rates. Beyond accuracy, acoustic analyses show that phonological proximity drives cross-lingual alignment, with Hindi–Marathi exhibiting the strongest correlations in duration, energy, and pitch. Spectrograms reveal a persistent gap: generated speech shows spectral sparsity and reduced richness compared to natural speech.

Future work will expand VAANI to more low-resource languages and domains, while focusing on acoustically grounded models that better handle code-switching and capture natural prosody.

Limitations

While this work makes progress in spontaneous speech-to-speech translation for Indian languages, the dataset currently comprises approximately 10 hours of parallel spontaneous speech, which, though useful for benchmarking, remains small compared to large-scale corpora such as VoxPopuli or FLEURS and may limit the training of high-capacity models. Our focus on Gujarati, Hindi, and Marathi provides valuable insights for this language group but does not capture the full linguistic diversity of India, leaving out Dravidian and many other Indo-Aryan languages. The data collection is also domain-specific, with recordings primarily from conversations, interviews, and discussions; as a result, the models trained or evaluated on this dataset may not generalize well to other domains such as broadcast news, medical dialogues, or casual telephone speech.

Ethics Statement

In developing and releasing this dataset, we were mindful of ethical considerations around data collection, privacy, and potential misuse. All speech recordings were either publicly available or collected with informed consent, and we ensured that no personally identifiable information or sensitive metadata was included, thereby reducing the risk of re-identification.

By focusing on underrepresented Indian languages, our goal is to advance equitable access to speech technologies; however, we also acknowledge that disparities in computational resources may limit who can fully benefit from these contributions.

References

Azza Abdel Monem, Khaled Shaalan, Ahmed Rafea, and Hoda Baraka. 2008. Generating arabic text in multilingual speech-to-speech machine translation framework. *Machine translation*, 22(4):205–258.

Rozin Majeed Abdullah, Siddeeq Y Ameen, Dindar Mikaeel Ahmed, Shakir Fattah Kak, Hajar Maseeh Yasin, Ibrahim Mahmood Ibrahim, Awder Mohammed Ahmed, Zryan Najat Rashid, Naaman Omar, and Azar Abid Salih. 2021. Paralinguistic speech processing: An overview. *Asian Journal of Research in Computer Science*, pages 34–46.

Penny P. Anderson and Emily Schrag Fenichel. 1989. Serving culturally diverse families of infants and toddlers with disabilities. ERIC Document ED318174.

Wim Codenie, Koen De Hondt, Patrick Steyaert, and Arlette Vercammen. 1997. From custom applications to domain-specific frameworks. *Communications of the ACM*, 40(10):70–77.

Sandeep Dhawan. 2022. Speech to speech translation: Challenges and future. *International Journal of Computer Applications Technology and Research*, 11(03):36–55.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, and 1 others. 2026. Waxal: A large-scale multilingual african language speech corpus. *arXiv preprint arXiv:2602.02734*.

Thierry Dutoit. 1997. High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, 17(1):25–36.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. 2022. Cascade or direct speech translation? a case study. *Applied Sciences*, 12(3):1097.

Steven Greenberg. 1997. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32.

Alessandro Gregori. 2021. Automatic speech recognition (asr) and nmt for interlingual and intralingual communication: Speech to text technology for live subtitling and accessibility.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Promptttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Deepak Gupta. 2022. Learning to answer multilingual and code-mixed questions. *arXiv preprint arXiv:2211.07522*.

Mohammad Reza Hasanabadi. 2023. An overview of text-to-speech systems and media applications. *arXiv preprint arXiv:2310.14301*.

Helenca Duxans i Barrobes. 2006. Voice conversion applied to text-to-speech systems. *Universitat Politecnica de Catalunya*.

Nwakanma Ifeanyi, Oluigbo Ikenna, and Okpala Izunna. 2014. Text–to–speech synthesis (tts). *International Journal of Research in Information Technology*, 2(5):154–163.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 13 indian languages. *arXiv preprint arXiv:2411.04699*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. Cvss corpus and massively multilingual speech-to-speech translation. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6691–6703.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2024. Textless unit-to-unit training for many-to-many multilingual speech-to-speech translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. 2023. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications*, 82(10):15171–15197.

Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang. 2024. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1557–1575.

Dominik Macháček. 2024. *Multi-Source Simultaneous Speech Translation*. Ph.D. thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

Vandan Mujadia and Dipti Misra Sharma. 2023. Towards speech to speech machine translation focusing on indian languages. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics: System demonstrations*, pages 161–168.

Marija Omazić and Martina Lekić. 2021. Assessing speech-to-speech translation quality: Case study of the ila s2s app. *Hieronymus: Journal of Translation Studies and Terminology*, (8):1–26.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing relies on linguistics. *Computational Linguistics*, 51(3):1009–1032.

Saadin Oyucu. 2023. A novel end-to-end turkish text-to-speech (tts) system via deep learning. *Electronics*, 12(8):1900.

Alejandro Manuel Pérez González de Martos. 2022. Deep neural networks for automatic speech-to-speech translation of open educational resources.

Rene Pilz and Johannes Schneider. 2025. Using phonemes in cascaded s2s translation pipeline. *arXiv preprint arXiv:2504.16234*.

John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. 2006. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1099–1108.

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.

Mr Mohit Saini. 2025. Dialect and standard language dynamics: analyzing code-switching and language adaptation in multilingual contexts. *GAP GYAN*, Volume VIII, Issue I.

Prithiv Sakthi. 2025. Common-voice-gender-detection (revision 0bca216). HuggingFace 10.57967/hf/5684.

Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2025. Towards building large scale datasets and state-of-the-art automatic speech translation systems for 14 indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32945–32966.

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

D Sasirekha and E Chandra. 2012. Text to speech: a simple tutorial. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):275–278.

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.

Thomas Rishi Sherborne. 2024. *Modelling cross-lingual transfer for semantic parsing*. Ph.D. thesis, University of Edinburgh.

Sanmay Sood, Siddharth Rajput, and Md Shad Akhtar. 2025. Speech-to-speech machine translation for dialectal variations of hindi. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 54–65.

Alistair Sutcliffe and Neil Maiden. 2002. The domain theory for requirements engineering. *IEEE Transactions on Software Engineering*, 24(3):174–196.

Isabel Trancoso, Francisco Teixeira, Catarina Botelho, and Alberto Abad. 2023. Treating speech as personally identifiable information and its impact in machine translation. In *Towards responsible machine translation: Ethical and legal considerations in machine translation*, pages 215–233. Springer.

Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, and Supriya Agrawal. 2018. Speech to text and text to speech recognition systems-areview. *IOSR J. Comput. Eng*, 20(2):36–43.

A. Waibel. 1988. *Prosody and Speech Recognition*. Research notes in artificial intelligence. Elsevier Science.

Rosa Yeh. 2014. *Effective strategies for using text-to-speech, speech-to-text, and machine-translation technology for teaching Chinese: A multiple-case study*. Northcentral University.

Jinyi Zhang, Ke Su, Haowei Li, Jiannan Mao, Ye Tian, Feng Wen, Chong Guo, and Tadahiro Matsumoto. 2024. Neural machine translation for low-resource languages from a chinese-centric perspective: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(6):1–60.

A Extended Related Work

A.1 Cascaded vs. Direct Speech Translation

Traditional approaches to speech translation rely on cascaded pipelines comprising Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) modules (Pérez González de Martos, 2022). While benefiting from the maturity of these individual components, cascaded architectures inherently suffer from error propagation, where inaccuracies in early stages amplify throughout the pipeline. Furthermore, they frequently fail to capture the rich paralinguistic features of spoken language, such as prosody, disfluencies, and informal speech patterns (Waibel, 1988). These limitations are particularly detrimental in Indian linguistic contexts, which are characterized by extensive code-switching, dialectal diversity, and regional phonetic variations (Saini, 2025).

To overcome the compounding errors of cascaded systems, direct Speech-to-Speech (S2S) translation models have emerged. These architectures map source speech directly to target speech without relying on intermediate textual representations (Ma et al., 2024; Kakwani et al., 2020). Although direct models promise end-to-end optimization and richer prosodic transfer, their development is severely bottlenecked by the scarcity of large-scale, high-quality parallel speech corpora, particularly for low-resource Indian languages (Ramesh et al., 2022).

A.2 Existing Datasets and Benchmarks

A review of the current landscape reveals a critical shortage of resources tailored for end-to-end S2S translation. For instance, BhasaAnuvaad (Jain et al., 2024) represents the largest publicly available dataset for Automatic Speech Translation in Indian languages, encompassing over 44,400 hours of speech across 13 languages. However, its primary focus, alongside subsets like SeamlessAlign (Sankar et al., 2025) and related corpora such as IndicTEDST (Sethiya et al., 2024), remains strictly on Speech-to-Text (S2T) translation. Consequently, these datasets do not adequately support the training or evaluation of direct S2S architectures.

B Spectrogram Analysis

Natural human speech is a complex, continuous signal characterized by its acoustic richness and inherent variability. The brighter, filled-in appearance is due to the presence of dense energy across the frequency spectrum. While high-energy regions like formants (the horizontal bright bands) are the most prominent features, the areas between them are not entirely silent. Real speech contains continuous, low-level energy, or noise, generated by several factors: slight aperiodicity in vocal fold vibration (vocal jitter), breath sounds, and micro-tremors that slightly blur the harmonic lines. These subtle components effectively raise the noise floor, preventing large swaths of the spectrogram from reaching the near-zero amplitude levels that register as pure black or dark spots. The rapid, fluid transitions between different speech sounds further contribute to this continuous, brighter texture.

In contrast, artificially generated speech often exhibits spectral sparsity which results in the prominent dark spots. This is particularly true for older or simpler synthesis models, but even sophisticated neural systems can sometimes be over-optimized for clarity. These synthetic systems generate sound based on idealized mathematical models, producing perfect, isolated harmonics exactly where they should be. Because they lack the natural, continuous background noise and the messy micro-variations of a human voice, the regions between the intended harmonics remain truly devoid of energy. These spots of near-zero amplitude appear as sharp, dark, empty gaps in the spectrogram. The resulting image is a stripped-down, simpler representation of the signal, showing a much sharper distinction between the high-energy (bright) formants and the zero-energy (dark) background.

C Evaluation Metrics

BLEU (Bilingual Evaluation Understudy). This is one of the earliest metrics, which evaluates the precision of n-gram overlaps between a system-generated text and reference texts, with a brevity penalty to avoid overly short outputs. Although it is effective in measuring surface-level accuracy, it may not capture deeper semantic meaning and fluency.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE is recall-focused and measures how much of the reference content is covered by the generated text, using variants such as ROUGE-N for n-gram overlap, ROUGE-L for longest common subsequence, and ROUGE-S for skip-bigrams.

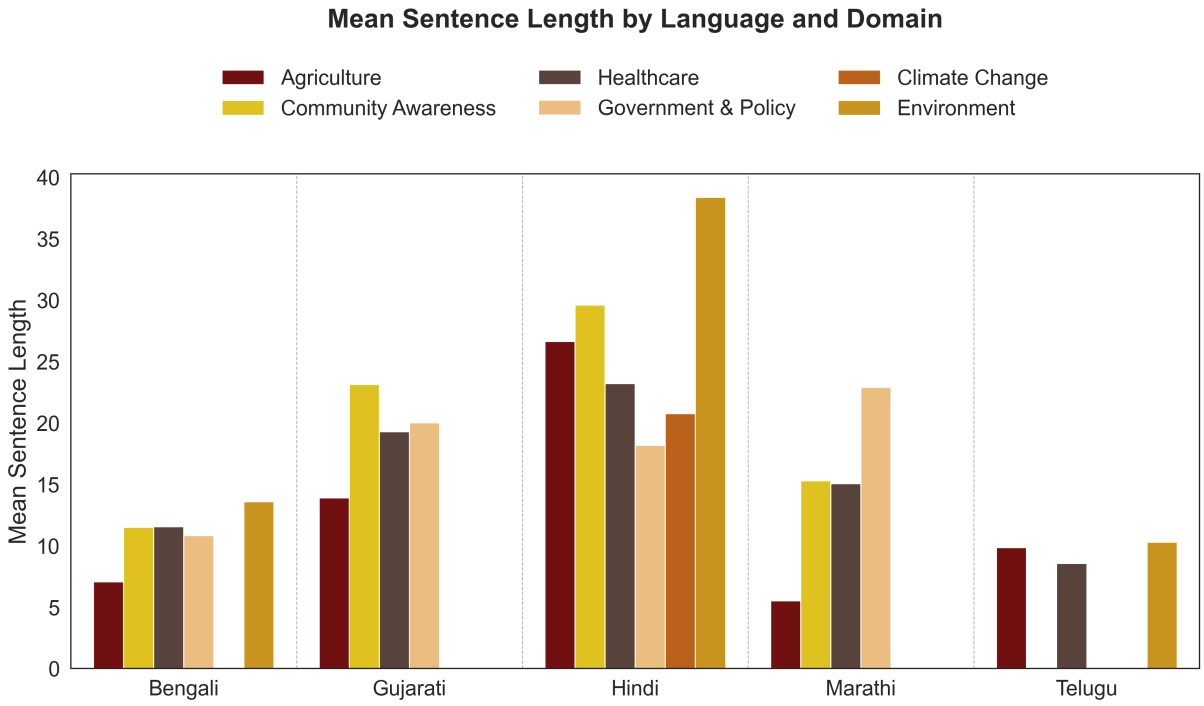


Figure 5: Domain vs Language: Mean Sentence Length.

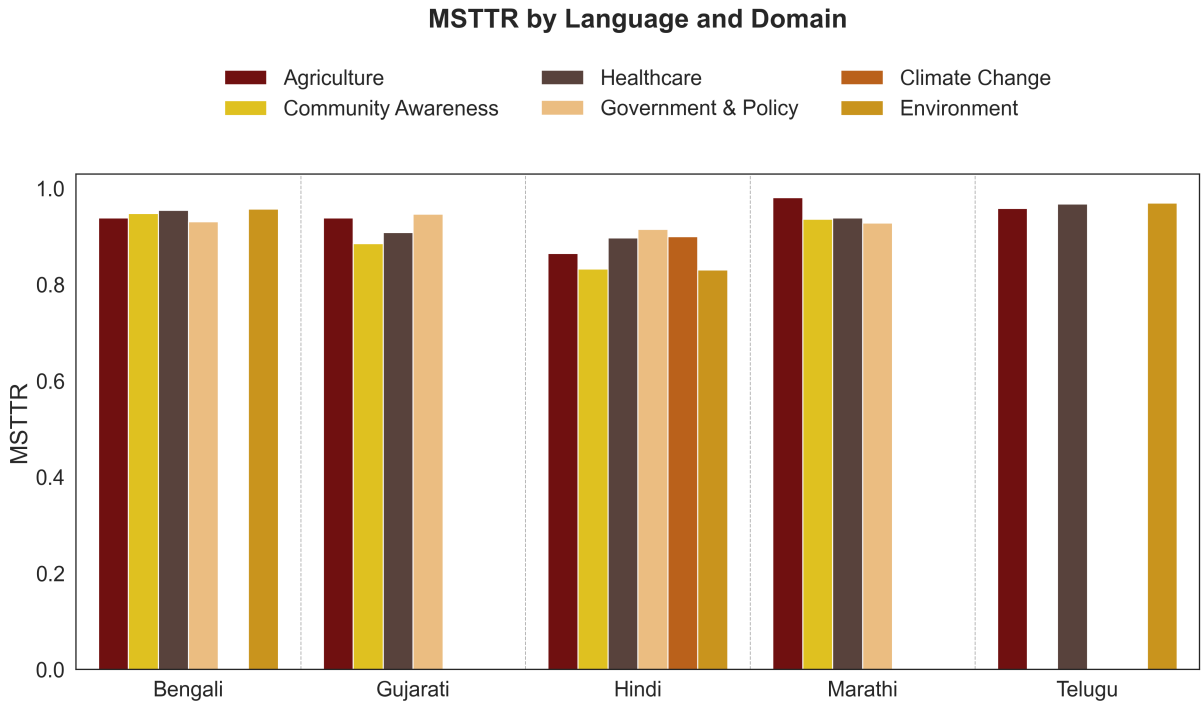


Figure 6: Domain vs Language: MSTTR.

METEOR (Metric for Evaluation of Translation with Explicit ORdering). METEOR improves upon BLEU by aligning candidate and reference words not only through exact matches but also via stemming, synonyms, and paraphrases, while incorporating penalties for incorrect word order. It balances precision and recall and tends to correlate more strongly with human judgment.

BERTScore. BERTScore takes advantage of contextual embeddings from pretrained language models such as BERT to measure semantic similarity between candidate and reference texts. Instead of relying only on exact word overlap, it compares

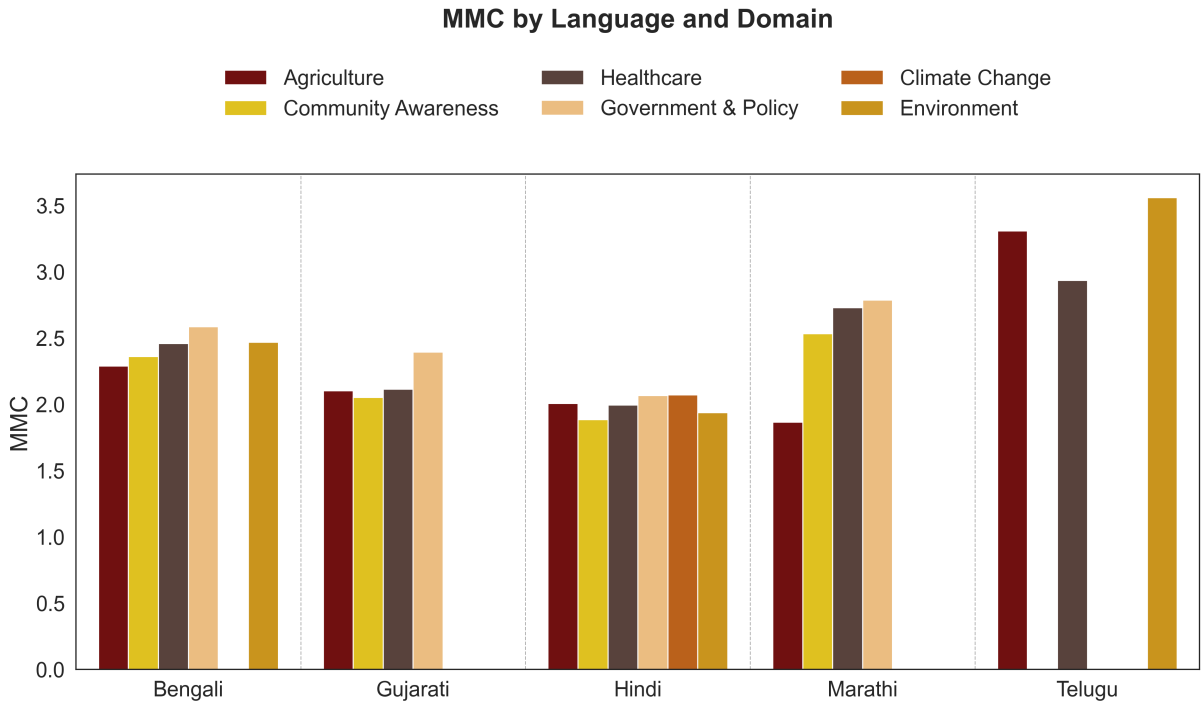


Figure 7: Domain vs Language: Mean Morphological Complexity (MMC).

token embeddings using cosine similarity, making it better at capturing meaning and nuance in text.

Mel Cepstral Distortion (MCD). A spectral distance measure that quantifies the difference between two speech signals in the mel-cepstral domain. Lower values indicate closer similarity in timbre and spectral envelope between generated and reference speech.

Pitch correlation. A measure of how well the pitch contour (fundamental frequency, F0) of the generated speech aligns with the reference speech. High correlation means the system preserves prosody and intonation patterns.

Duration Correlation. Evaluates the similarity in timing and rhythm between generated and reference speech by correlating phoneme or word durations. High correlation indicates better temporal alignment and naturalness.

Energy Correlation. Measures how closely the energy (loudness/intensity) patterns of generated speech match those of the reference. High correlation means the output preserves stress and emphasis.

Speaker Embedding Cosine Similarity. Compares speaker identity by calculating the cosine similarity between speaker embeddings (vector representations). A higher value indicates the generated speech is more similar in voice characteristics to the reference speaker.

Structural Similarity Index (SSIM). Originally designed for images, but adapted here for speech spectrograms. It measures perceived similarity by considering luminance, contrast, and structural patterns. Higher SSIM indicates the generated spectrogram is more similar to the reference.

D Statistical Analysis of Linguistic Metrics

To ensure the robustness of our linguistic analysis, we evaluated the interdependencies of our chosen metrics and quantified the primary drivers of linguistic variation within our dataset.

D.1 Metric Interdependencies and Independence

Table 4 details the Spearman rank correlation coefficients between the five evaluated linguistic metrics. The analysis reveals several key structural relationships:

Syntactic Overlap. Mean Sentence Length (MSL) and Average Weighted Sentence Complexity (AWSC) exhibit a near-perfect positive correlation ($r = 0.99$). This indicates that within these

Table 6: **Speech-based objective metrics for cross-lingual transfer across five languages**. Each block corresponds to a base (training) language, and columns denote target (evaluation) languages.

Metric	**Target Languages**			
Base Language: Hindi	**Gujarati**	**Marathi**	**Telugu**	**Bengali**
Mel Cepstral Distortion	2.96	3.02	2.92	3.56
Pitch Correlation	0.6305	0.7577	0.6365	0.7594
Duration Correlation	0.6604	0.7542	0.6608	0.7549
Energy Correlation	0.6551	0.7509	0.6558	0.7559
Speaker Embedding Cosine Similarity	0.6566	0.7671	0.6569	0.7670
Structural Similarity Index	0.4755	0.4011	0.6578	0.7670
Base Language: Gujarati	**Hindi**	**Marathi**	**Telugu**	**Bengali**
Mel Cepstral Distortion	2.93	3.04	2.98	3.06
Pitch Correlation	0.6255	0.7477	0.6259	0.7476
Duration Correlation	0.6576	0.7442	0.6250	0.7211
Energy Correlation	0.6455	0.7503	0.6250	0.7470
Speaker Embedding Cosine Similarity	0.6411	0.7666	0.6244	0.7470
Base Language: Marathi	**Hindi**	**Gujarati**	**Telugu**	**Bengali**
Mel Cepstral Distortion	2.90	2.94	2.95	2.99
Pitch Correlation	0.6065	0.7486	0.6242	0.7444
Duration Correlation	0.6233	0.7455	0.6225	0.7470
Energy Correlation	0.6315	0.7565	0.6231	0.7470
Speaker Embedding Cosine Similarity	0.6233	0.7766	0.6233	0.7470
Structural Similarity Index	0.4996	0.4211	0.4896	0.4463
Base Language: Telugu	**Hindi**	**Gujarati**	**Marathi**	**Bengali**
Mel Cepstral Distortion	2.83	2.90	2.77	2.95
Pitch Correlation	0.6024	0.7480	0.6244	0.7470
Duration Correlation	0.6223	0.7255	0.6244	0.7470
Energy Correlation	0.6215	0.7465	0.6455	0.7470
Speaker Embedding Cosine Similarity	0.6231	0.7666	0.6477	0.7666
Structural Similarity Index	0.4977	0.4202	0.4779	0.4443
Base Language: Bengali	**Hindi**	**Gujarati**	**Marathi**	**Telugu**
Mel Cepstral Distortion	2.80	2.91	2.80	2.95
Pitch Correlation	0.6022	0.7444	0.6287	0.7332
Duration Correlation	0.6213	0.7355	0.6286	0.7470
Energy Correlation	0.6244	0.7526	0.6289	0.7455
Speaker Embedding Cosine Similarity	0.6126	0.7563	0.6290	0.7470
Structural Similarity Index	0.4996	0.4211	0.4999	0.4418

Indic languages, sentence length is the primary driver of the composite AWSC score.

Lexical Density. MSL demonstrates a strong negative correlation with the Mean Segmental Type-Token Ratio (MSTTR) at $r = -0.67$. This reflects a common linguistic pattern where lexical diversity naturally decreases as utterance length increases.

Morphological Independence. Mean Morphological Complexity (MMC) operates almost entirely independently of the other metrics. It shows negligible correlations with MSL ($r = 0.03$), AWSC ($r = 0.09$), and Mean Clause Depth ($r = 0.09$). This confirms that MMC successfully captures a distinct morphological dimension–evaluating word formation and suffix density–that is structurally separate from syntactic length or clause depth.

D.2 Drivers of Linguistic Variation

To understand the sources of variance within the corpus, we conducted an Omnibus Kruskal-Wallis analysis, as detailed in Table 5. The results confirm that both the spoken language and the topical domain are highly significant drivers of linguistic variation, yielding p-values of < 0.001 across all

Table 7: **Evaluation results for cross-lingual sentence generation.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Marathi	Gujarati	**0.6725**	**0.6321**	0.5679	45.67	**0.6228**	0.6256	0.6466	0.0267	0.0459	**0.0170**
Marathi	Hindi	0.6303	0.5788	0.5771	48.90	0.6026	**0.6456**	**0.6688**	0.0272	0.0455	0.0175
Marathi	Telugu	0.6203	0.5925	0.5966	**48.97**	0.6128	0.6233	0.6480	**0.0264**	**0.0450**	0.0171
Marathi	Bengali	0.6209	0.5825	**0.5976**	48.22	0.6055	0.6433	0.6470	0.0268	0.0459	0.0170
Hindi	Marathi	**0.6712**	**0.6227**	0.5622	**47.89**	**0.6223**	0.6270	0.6482	**0.0270**	0.0555	0.0179
Hindi	Gujarati	0.6555	0.6218	0.5744	45.12	0.6012	0.6272	0.6473	0.0278	0.0452	**0.0171**
Hindi	Telugu	0.6408	0.5957	0.5988	46.90	0.6066	0.6280	0.6466	0.0274	0.0355	0.0275
Hindi	Bengali	0.6455	0.5984	**0.5996**	46.98	0.6078	**0.6290**	**0.6488**	0.0352	**0.0255**	0.0178
Gujarati	Marathi	**0.6771**	0.6184	0.5506	**46.23**	0.6241	**0.6288**	0.6368	0.0275	0.0451	**0.0170**
Gujarati	Hindi	0.6308	**0.6219**	0.5561	45.12	0.6012	0.6285	0.6478	0.0274	0.0459	0.0171
Gujarati	Telugu	0.6514	0.5976	**0.5996**	46.96	0.6080	0.6283	**0.6480**	0.0274	0.0458	0.0172
Gujarati	Bengali	0.6517	0.5906	0.5991	45.96	**0.6512**	0.6288	0.6416	**0.0268**	**0.0441**	0.0170
Telugu	Marathi	**0.6760**	0.6140	0.5502	46.20	**0.6220**	**0.6283**	0.6488	0.0274	**0.0345**	**0.0170**
Telugu	Hindi	0.6204	**0.6206**	0.5446	45.18	0.6019	0.6241	0.6440	0.0274	0.0451	0.0170
Telugu	Gujarati	0.6588	0.5986	0.5990	**46.92**	0.6040	0.6278	**0.6480**	0.0270	0.0459	0.0171
Telugu	Bengali	0.6378	0.5980	**0.5996**	45.96	0.6080	0.6284	0.6298	**0.0266**	0.0357	0.0162
Bengali	Marathi	**0.6750**	0.6130	0.5455	46.10	**0.6120**	0.6183	**0.6433**	0.0277	**0.0351**	0.0270
Bengali	Hindi	0.6133	**0.6208**	0.5449	45.05	0.6054	**0.6223**	0.6240	**0.0263**	0.0450	0.0178
Bengali	Gujarati	0.6388	0.5922	0.5970	**46.42**	0.6020	0.6218	0.6280	0.0270	0.0450	0.0179
Bengali	Telugu	0.6322	0.5956	**0.5989**	45.83	0.6056	0.6266	0.6290	0.0270	0.0447	**0.0168**

evaluated metrics.

The Language Effect. The specific language spoken exerts the most profound statistical effect on morphological complexity (MMC, H-stat = 1845.76) and sentence length (MSL, H-stat = 1088.88). This highlights the deep inherent structural and grammatical differences between the five evaluated languages.

The Domain Effect. The topical context also heavily influences structural complexity, particularly impacting morphological complexity (MMC, H-stat = 590.06) and overall sentence length/complexity (MSL and AWSC, H-stats $\approx$ 475). This suggests that technical or specialized topics inherently demand more complex word formations and longer utterances compared to casual domains.

Table 8: **Evaluation results for sentence-level generation using *whisper-large-v3*.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Gujarati	Gujarati	0.8521	0.8124	0.8036	81.64	0.8421	**0.8556**	0.8744	0.0281	**0.0181**	0.0462
Gujarati	Hindi	**0.8605**	**0.8322**	**0.8254**	**82.65**	**0.8551**	0.8551	0.8748	0.0330	0.0282	0.0360
Gujarati	Marathi	0.8101	0.7971	0.7945	81.21	0.8225	0.8456	**0.8788**	0.0160	0.0280	0.0380
Gujarati	Telugu	0.8005	0.7904	0.7921	81.04	0.8220	0.8512	0.8777	0.0155	0.0270	0.0265
Gujarati	Bengali	0.8025	0.7909	0.7920	81.09	0.8230	0.8518	0.8785	**0.0145**	0.0370	**0.0262**
Hindi	Hindi	**0.8773**	**0.8563**	0.8328	**84.60**	0.8540	**0.8583**	**0.8770**	0.0282	**0.0180**	0.0481
Hindi	Gujarati	0.8534	0.8424	**0.8329**	82.22	**0.8555**	0.8580	0.8760	0.0340	0.0263	**0.0350**
Hindi	Marathi	0.8401	0.8255	0.7834	82.45	0.8511	0.8580	0.8767	0.0180	0.0275	0.0381
Hindi	Telugu	0.8206	0.8152	0.7755	81.22	0.8377	0.8572	0.8767	**0.0150**	0.0225	0.0360
Hindi	Bengali	0.8201	0.8158	0.7758	81.28	0.8367	0.8570	0.8758	0.0154	0.0220	0.0366
Marathi	Marathi	**0.8864**	**0.8559**	**0.8423**	**84.12**	0.8234	0.8575	0.8754	0.0280	**0.0162**	0.0437
Marathi	Gujarati	0.8520	0.8417	0.8334	82.55	**0.8498**	0.8552	0.8777	0.0310	0.0262	0.0360
Marathi	Hindi	0.8408	0.8280	0.7886	82.70	0.8477	0.8580	**0.8788**	0.0160	0.0281	0.0394
Marathi	Telugu	0.8206	0.8240	0.7875	82.30	0.8215	0.8583	0.8784	**0.0140**	0.0250	0.0330
Marathi	Bengali	0.8203	0.8247	0.7879	82.36	0.8218	**0.8580**	0.8788	0.0143	0.0256	**0.0230**
Telugu	Telugu	**0.8655**	**0.8486**	**0.8401**	**83.68**	0.8239	**0.8475**	0.8554	0.0260	**0.0160**	0.0427
Telugu	Gujarati	0.8466	0.8377	0.8255	81.67	0.8215	0.8372	0.8758	0.0256	0.0260	0.0350
Telugu	Hindi	0.8355	0.8240	0.7880	82.40	**0.8430**	0.8579	0.8814	0.0155	0.0280	0.0390
Telugu	Marathi	0.8255	0.8230	0.7845	82.20	0.8330	0.8375	0.8850	**0.0127**	0.0240	0.0320
Telugu	Bengali	0.8250	0.8237	0.7849	82.40	0.8350	0.8372	**0.8861**	0.0128	0.0230	**0.0319**
Bengali	Bengali	0.8413	**0.8422**	**0.8255**	**83.13**	0.8222	0.8423	0.8514	0.0265	**0.0144**	0.0421
Bengali	Telugu	**0.8414**	0.8244	0.8212	81.24	0.8206	0.8313	0.8715	0.0250	0.0166	0.0423
Bengali	Hindi	0.8044	0.8051	0.7655	82.04	**0.8413**	**0.8522**	0.8655	0.0250	0.0240	**0.0250**
Bengali	Marathi	0.8203	0.8130	0.7841	82.10	0.8310	0.8345	**0.8820**	0.0145	0.0270	0.0370
Bengali	Gujarati	0.8212	0.8210	0.7811	82.18	0.8317	0.8278	0.8640	**0.0123**	0.0230	0.0310

Table 9: **Evaluation results for sentence-level generation using *whisper-large-v2*.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Hindi	Hindi	0.8455	**0.8265**	0.8034	81.92	0.7955	0.8571	0.8758	0.0281	0.0181	0.0462
Hindi	Gujarati	0.8496	0.8056	**0.8287**	81.67	**0.7999**	0.8567	0.8743	**0.0074**	**0.0050**	0.0280
Hindi	Marathi	**0.8594**	0.8237	0.8075	**83.66**	0.7997	0.8560	0.8655	0.0330	0.0282	0.0360
Hindi	Telugu	0.8506	0.8137	0.8022	81.63	0.7944	**0.8573**	**0.8759**	0.0160	0.0280	0.0380
Hindi	Bengali	0.8509	0.8127	0.8010	81.41	0.7925	0.8570	0.8750	0.0155	0.0270	**0.0265**
Gujarati	Gujarati	0.8450	**0.8261**	0.8024	81.84	0.7922	**0.8275**	0.8351	0.0282	0.0180	0.0481
Gujarati	Hindi	0.8492	0.8051	**0.8282**	81.62	**0.7993**	0.8075	**0.8655**	**0.0110**	**0.0062**	**0.0295**
Gujarati	Marathi	**0.8590**	0.8232	0.8070	**83.60**	0.7992	0.8175	0.8254	0.0340	0.0263	0.0350
Gujarati	Telugu	0.8580	0.8230	0.8055	82.60	0.7844	0.8075	0.8254	0.0180	0.0275	0.0381
Gujarati	Bengali	0.8570	0.8220	0.8051	82.15	0.7840	0.8065	0.8250	0.0150	0.0225	0.0360
Marathi	Marathi	0.8456	**0.8255**	0.8052	81.56	0.7877	0.8075	0.8254	0.0280	**0.0162**	0.0437
Marathi	Gujarati	0.8467	0.7990	**0.8355**	82.88	0.7788	0.8270	0.8154	0.0310	0.0262	0.0360
Marathi	Hindi	**0.8584**	0.8230	0.8176	**84.47**	**0.7897**	**0.8560**	**0.8750**	0.0160	0.0281	0.0394
Marathi	Telugu	0.8580	0.8220	0.8125	84.23	0.7825	0.8560	0.8733	**0.0140**	0.0250	0.0330
Marathi	Bengali	0.8567	0.8213	0.8122	84.26	0.7825	0.8558	0.8738	0.0143	0.0256	**0.0230**
Telugu	Telugu	**0.8455**	**0.8480**	**0.8315**	**83.22**	0.8213	0.8215	**0.8750**	0.0260	0.0160	0.0427
Telugu	Gujarati	0.8366	0.8315	0.8222	81.22	0.8204	0.8275	0.8733	0.0820	**0.0062**	**0.0268**
Telugu	Hindi	0.8323	0.8220	0.7870	82.20	**0.8410**	0.8170	0.8252	0.0155	0.0280	0.0390
Telugu	Marathi	0.8245	0.8220	0.7811	82.04	0.8310	0.8372	0.8744	**0.0127**	0.0240	0.0320
Telugu	Bengali	0.8246	0.8227	0.7810	82.01	0.8347	**0.8378**	0.8749	0.0128	0.0230	0.0319
Bengali	Bengali	**0.8450**	0.8422	**0.8310**	82.22	0.8210	0.8210	**0.8740**	0.0265	0.0144	0.0421
Bengali	Telugu	0.8450	**0.8460**	0.8306	**83.20**	0.8113	0.8210	0.8740	0.0250	0.0166	0.0423
Bengali	Gujarati	0.8360	0.8310	0.8220	81.25	0.8200	0.8270	0.8728	0.0820	**0.0062**	**0.0268**
Bengali	Hindi	0.8320	0.8224	0.7875	82.10	**0.8411**	0.8172	0.8052	0.0145	0.0270	0.0370
Bengali	Marathi	0.8240	0.8210	0.7810	82.01	0.8303	**0.8370**	0.8740	**0.0123**	0.0230	0.0310

Table 10: **Evaluation results for sentence-level generation using *seamless-4T-large-v2*.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Marathi	Marathi	0.8103	0.7955	0.7855	80.04	0.7312	**0.7866**	0.7951	0.0272	0.0455	**0.0175**
Marathi	Gujarati	**0.8223**	0.7865	**0.8113**	80.11	0.7312	0.7880	**0.7976**	0.0330	**0.0360**	0.0282
Marathi	Hindi	0.8207	0.8021	0.7841	**81.03**	**0.7495**	0.7812	0.7910	0.0160	0.0380	0.0281
Marathi	Telugu	0.8201	0.7883	0.7818	80.09	0.7426	0.7830	0.7921	0.0160	0.0380	0.0281
Marathi	Bengali	0.8168	**0.8110**	0.7802	80.41	0.7370	0.7890	0.7972	**0.0146**	0.0370	0.0272
Gujarati	Gujarati	0.8158	0.8034	0.7756	80.02	0.7233	**0.7840**	0.7851	0.0282	0.0481	**0.0182**
Gujarati	Marathi	**0.8204**	0.7812	**0.8105**	80.04	0.7305	0.7710	0.7750	0.0340	**0.0350**	0.0256
Gujarati	Hindi	0.8202	0.8002	0.7805	81.01	0.7492	0.7960	0.7750	0.0176	0.0377	0.0270
Gujarati	Telugu	0.8166	0.8001	0.7701	**81.08**	0.7344	0.7850	0.7976	0.0156	0.0370	0.0260
Gujarati	Bengali	0.8169	**0.8115**	0.7808	80.47	0.7375	0.7896	**0.7977**	**0.0143**	0.0384	0.0230
Hindi	Hindi	0.8326	0.8023	0.7901	80.22	0.7312	0.7845	0.7650	0.0273	0.0437	**0.0155**
Hindi	Gujarati	0.8306	0.7902	**0.8123**	80.11	0.7237	**0.7911**	0.7851	0.0316	**0.0360**	0.0257
Hindi	Marathi	0.8302	0.8112	0.7822	**81.02**	0.7421	0.7880	0.7950	0.0316	0.0360	0.0257
Hindi	Telugu	0.8166	0.8110	0.7806	80.02	0.7358	0.7888	0.7952	**0.0156**	0.0388	0.0280
Hindi	Bengali	0.8176	**0.8118**	0.7845	80.78	0.7380	0.7899	**0.7986**	0.0165	0.0389	0.0286
Telugu	Telugu	0.8633	0.8266	**0.8266**	83.15	0.8215	0.7860	0.7958	0.0215	**0.0358**	0.0260
Telugu	Gujarati	0.8413	**0.8333**	0.8215	81.25	0.8166	0.8066	0.8351	0.0355	0.0370	0.0276
Telugu	Hindi	0.8320	0.8220	0.7860	82.20	**0.8410**	**0.8666**	**0.8951**	0.0410	0.0551	**0.1666**
Telugu	Marathi	0.8220	0.8210	0.7822	81.20	0.8258	0.8466	0.8551	**0.0184**	0.0431	0.0294
Telugu	Bengali	0.8227	0.8218	0.7847	81.46	0.8455	0.8488	0.8587	0.0189	0.0430	0.0291
Bengali	Bengali	0.8425	0.8066	**0.8104**	81.22	0.8058	0.7566	0.7921	0.0282	0.0494	**0.0190**
Bengali	Telugu	0.8411	0.8031	0.8031	83.01	0.8044	0.7560	0.7752	0.0281	0.0480	0.0192
Bengali	Gujarati	0.8213	0.8120	0.8055	81.03	0.8044	0.8021	0.8125	0.0350	0.0373	0.0271
Bengali	Hindi	0.8112	**0.8121**	0.7655	80.33	0.8211	0.8311	**0.8751**	**0.0180**	0.0423	0.0291
Bengali	Marathi	0.8020	0.8010	0.7820	81.10	0.8122	**0.8266**	0.8151	0.0352	**0.0362**	0.0270

Table 11: **Evaluation results for sentence-level generation using *AudioPaLM-2*.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Marathi	Marathi	0.8236	0.8067	0.7922	80.83	0.7377	0.7868	0.7955	0.0282	0.0463	**0.0181**
Marathi	Gujarati	**0.8344**	0.7906	**0.8255**	80.77	0.7466	0.7854	**0.7958**	0.0333	**0.0366**	0.0285
Marathi	Hindi	0.8282	**0.8178**	0.7899	**82.03**	**0.7673**	0.7840	0.7941	0.0162	0.0384	0.0285
Marathi	Telugu	0.8280	0.8166	0.7834	81.26	0.7622	0.7860	0.7930	**0.0162**	0.0384	0.0285
Marathi	Bengali	0.8287	0.8160	0.7837	80.26	0.7648	**0.7889**	0.7885	0.0164	0.0380	0.0288
Gujarati	Gujarati	0.8255	**0.8166**	0.7877	81.55	0.7366	0.7862	0.7950	0.0284	0.0486	**0.0185**
Gujarati	Marathi	**0.8355**	0.7956	**0.8266**	**81.44**	0.7456	0.8066	0.7958	0.0345	**0.0358**	0.0265
Gujarati	Hindi	0.8277	0.8102	0.7908	82.66	**0.7597**	**0.8266**	0.7941	0.0182	0.0384	0.0276
Gujarati	Telugu	0.8250	0.8050	0.7883	82.16	0.7527	0.7870	**0.7975**	0.0176	0.0375	0.0270
Gujarati	Bengali	0.8278	0.8020	0.7889	82.67	0.7427	0.7890	0.7942	**0.0170**	**0.0274**	0.0277
Hindi	Hindi	0.8377	**0.8355**	0.7966	80.88	0.7366	0.7880	**0.7975**	0.0286	0.0482	**0.0189**
Hindi	Gujarati	0.8377	0.7968	**0.8188**	80.88	0.7298	0.7860	0.7949	0.0346	**0.0350**	0.0269
Hindi	Marathi	**0.8477**	0.8218	0.7966	**82.09**	**0.7566**	**0.7890**	0.7940	0.0180	0.0388	0.0271
Hindi	Telugu	0.8451	0.8203	0.7931	82.02	0.7522	0.7878	0.7945	**0.0175**	0.0370	0.0278
Hindi	Bengali	0.8450	0.8208	0.7927	82.09	0.7528	0.7855	0.7940	0.0175	**0.0275**	0.0279
Telugu	Telugu	0.8122	**0.8086**	0.8001	79.68	0.8039	0.7860	0.7941	0.0280	0.0440	**0.0160**
Telugu	Gujarati	**0.8166**	0.8077	**0.8201**	79.61	0.8017	0.7882	0.7977	0.0320	**0.0362**	0.0261
Telugu	Hindi	0.8105	0.8005	0.7644	80.15	**0.8216**	0.8066	**0.8124**	0.0152	0.0455	0.0289
Telugu	Marathi	0.8055	0.7885	0.7640	80.10	0.8106	0.8266	0.7988	0.0145	0.0459	0.0296
Telugu	Bengali	0.8058	0.7880	0.7630	**80.67**	0.8146	**0.8466**	0.7926	**0.0140**	0.0450	0.0267
Bengali	Bengali	0.8023	0.8022	0.8005	79.97	0.8066	0.7868	0.7949	0.0155	0.0460	0.0293
Bengali	Telugu	0.8128	**0.8080**	0.8000	79.60	0.8031	0.7865	0.7934	0.0288	0.0441	**0.0163**
Bengali	Gujarati	**0.8160**	0.8071	**0.8205**	79.66	0.8019	0.7880	0.7970	**0.0320**	**0.0360**	0.0260
Bengali	Hindi	0.8107	0.8002	0.7640	80.10	**0.8211**	0.8060	**0.8120**	0.0151	0.0462	0.0290
Bengali	Marathi	0.8050	0.7888	0.7630	**80.16**	0.8108	**0.8260**	0.7980	0.0150	**0.0468**	0.0297

Table 12: **Evaluation results for sentence-level generation using *Seamless-4T-large*.** Metrics reported are Rouge-2 F1 (**R2-F1**), Rouge-L F1 (**RL-F1**), METEOR (**ME**), BLEU (**BL**), BERTScore (**BS**), Word Error Rate (**WER**), Character Error Rate (**CER**), and Phoneme Error Rate (**PER**). Results are shown for each base language and its translated counterparts.

Base	Target	R2-F1	RL-F1	ME	BL	BS	IBS	MBS	WER	CER	PER
Marathi	Marathi	0.8056	0.7922	0.7811	79.04	0.7244	0.7750	0.7846	0.0280	**0.0180**	0.0466
Marathi	Gujarati	0.8145	0.7833	**0.8103**	80.02	0.7255	**0.7780**	0.7867	0.0323	0.0281	**0.0362**
Marathi	Hindi	**0.8201**	0.8005	0.7811	**81.01**	0.7422	0.7760	0.7880	**0.0151**	0.0273	0.0386
Marathi	Telugu	0.8106	0.8002	0.7803	80.01	0.7406	0.7760	**0.8146**	0.0320	0.0286	0.0367
Marathi	Bengali	0.8109	**0.8008**	0.7807	80.08	**0.7434**	0.7777	0.8066	0.0158	0.0276	0.0388
Gujarati	Gujarati	0.8133	0.8022	0.7721	80.00	0.7212	0.7770	0.7867	0.0277	0.0180	0.0476
Gujarati	Marathi	0.8108	0.7803	**0.8055**	79.55	0.7244	0.7740	0.7826	0.0345	**0.0260**	**0.0350**
Gujarati	Hindi	0.8153	0.7945	0.7705	**81.00**	0.7422	0.7730	0.7988	0.0177	0.0270	0.0385
Gujarati	Telugu	0.8140	0.7922	0.7655	81.05	0.7311	0.7740	0.7821	**0.0171**	0.0276	0.0382
Gujarati	Bengali	**0.8156**	**0.8066**	0.7863	80.77	**0.7466**	**0.7798**	**0.8088**	0.0175	**0.0344**	0.0277
Hindi	Hindi	**0.8311**	**0.8004**	0.7844	79.55	0.7303	0.7730	0.7826	0.0286	**0.0166**	0.0447
Hindi	Gujarati	0.8306	0.7902	**0.8123**	80.11	0.7237	0.7850	0.7946	0.0324	0.0262	**0.0363**
Hindi	Marathi	0.8244	0.8024	0.7999	**81.00**	0.7411	**0.7950**	0.8046	0.0162	0.0277	0.0388
Hindi	Telugu	0.8150	0.7921	0.7655	80.07	0.7404	0.7850	0.8146	0.0170	0.0262	0.0380
Hindi	Bengali	0.8187	0.7956	0.7680	80.77	**0.7564**	0.7889	**0.8188**	**0.0155**	0.0234	0.0389
Telugu	Telugu	**0.8265**	0.8005	0.8001	**80.61**	0.7839	**0.7957**	**0.8246**	**0.0158**	**0.0306**	**0.0469**
Telugu	Gujarati	0.8166	0.8051	0.8051	78.61	0.7810	0.7744	0.7848	0.0280	0.0163	0.0445
Telugu	Hindi	0.8016	**0.8066**	0.7612	82.20	0.8120	0.7850	0.7849	0.0344	0.0285	0.0384
Telugu	Marathi	0.8055	0.7990	0.7622	80.10	0.8106	0.7850	0.7994	0.0167	0.0298	0.0476
Telugu	Bengali	0.8089	0.7997	**0.7655**	80.50	**0.8156**	0.7878	0.7998	0.0169	0.0292	0.0470
Bengali	Bengali	**0.8260**	0.8008	0.8005	80.67	0.7832	0.7950	0.8240	**0.0150**	0.0300	0.0463
Bengali	Telugu	0.8260	0.8001	0.8007	80.66	0.7832	**0.7952**	**0.8240**	0.0151	0.0300	0.0462
Bengali	Gujarati	0.8160	0.8050	**0.8055**	78.67	0.7815	0.7740	0.7840	0.0270	**0.0160**	0.0443
Bengali	Hindi	0.8010	**0.8060**	0.7615	**82.27**	0.8020	0.7650	0.7644	0.0342	**0.0380**	0.0280
Bengali	Marathi	0.8022	0.7975	0.7605	80.04	**0.8102**	0.7840	0.7982	0.0163	0.0294	0.0478