Cascaded Modular or End-to-End? : An Investigation on Speech-to-Speech Translation Task for Dravidian Languages

Bhavana Nali
IIT Bhubaneswar, India
24ai06013@iitbbs.ac.in

Abhik Jana
IIT Bhubaneswar, India
abhikjana@iitbbs.ac.in

Abstract

This paper presents a study of speech-to-speech translation for low-resource Dravidian languages, focusing on Tamil, Telugu, and Kannada. We investigate the efficacy of the Cascaded Modular system with the End-to-End system in both zero-shot and fine-tuned settings. The Cascaded Modular approach combines an ASR Module (Whisper Large-v3 for English speech; IndicConformer for Dravidian speech), a Text-to-Text translation module (IndicTrans2), and a Speech synthesis module (Indic Parler-TTS), whereas SeamlessM4T-v2 is used as the End-to-End system. For parameter-efficient Low-Rank Adaptation (LoRA) fine-tuning to adapt the translation component to domain-specific datasets, we use FLEURS and Mann-ki-Baat (a subset of the BhasaAnuvaad dataset). Cascaded Modular systems achieve BLEU scores ranging from 3.17 to 18.96 in the zero-shot setting and 5.08 to 19.18 after fine-tuning, whereas the End-to-End model ranges from 3.02 to 15.72 in zero-shot settings across languages and 4.11 to 16.84 after fine-tuning. The results show that Cascaded Modular systems generally outperform the End-to-End model across most setups, though the margin varies across language pairs. Parameter-efficient fine-tuning yields notable improvements in translation quality and speech generation performance for low-resource Dravidian speech translation. The code repository is made publicly available[1].

1 Introduction

Speech-to-speech (S2S) translation enables direct spoken communication between speakers of different languages. Recent multilingual models such as Whisper (Radford et al., 2023) and SeamlessM4T (Seamless Communication Team, 2023) have advanced multilingual speech translation, yet most progress focuses on high-resource languages. Low-resource Dravidian languages such as Tamil, Telugu, and Kannada remain underexplored despite their linguistic diversity and limited parallel speech resources. Two main approaches exist for S2S translation: the *Cascaded Modular* approach, which combines ASR, Text-to-Text translation, and TTS modules, and the *End-to-End* approach, which directly maps source speech to target speech. While datasets such as FLEURS (Conneau et al., 2022) and BhasaAnuvaad (Sankar et al., 2025a) support multilingual research, systematic comparisons for Dravidian languages remain limited.

In this work, we study S2S translation for Tamil, Telugu, and Kannada under both paradigms. English-to-Dravidian translation uses Whisper Large-v3 (Radford et al., 2023), Dravidian-to-Dravidian uses IndicConformer (Bhogale et al., 2025), both pipelines use IndicTrans2 (Gala et al., 2023) and Indic Parler-TTS (Sankar et al., 2025b), and SeamlessM4T-v2 serves as the End-to-End baseline. We apply parameter-efficient LoRA fine-tuning (Hu et al., 2022) on FLEURS and Mann-ki-Baat (Sankar et al., 2025a). Cascaded models achieve BLEU of 3.17–18.96 (zero-shot) and 5.08–19.18 (fine-tuned); End-to-End ranges from 3.02–15.72 and 4.11–16.84, respectively. To the best of our knowledge, this is the first comprehensive investigation of S2S translation for these three Dravidian languages under both paradigms.

2 Methodology

We conduct a comparative study of two speech-to-speech translation pipelines: Cascaded Modular and End-to-End. Figure 1 illustrates the architecture of the pipelines used in this work.

Cascaded Modular Architecture: This framework comprises a three-stage pipeline consisting of Automatic Speech Recognition (ASR), Text-to-Text Translation (TT), and Text-to-Speech (TTS) modules.

[1] https://github.com/bhavananali/cascade-vs-e2e-dravidian

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 14–20

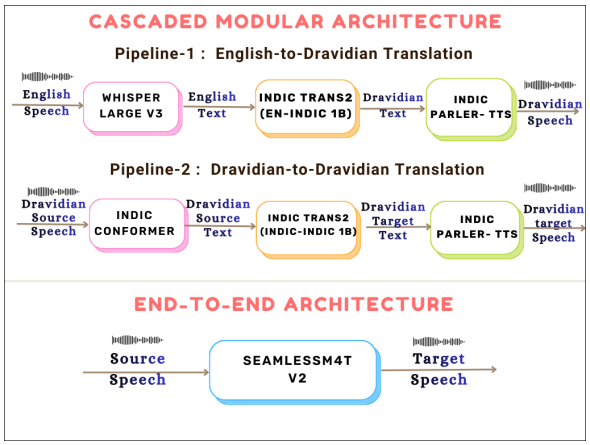

Figure 1: Comparison of speech-to-speech translation architectures. **Cascaded Modular Architecture** (Top): Pipeline 1 utilizes Whisper Large-v3 for English ASR, while Pipeline 2 uses Indic-Conformer for Dravidian ASR, both feeding into IndicTrans2 and finally Indic Parler-TTS. **End-to-End Architecture** (Bottom): Pipeline 3 uses SeamlessM4T v2 for direct speech-to-speech mapping.

Automatic Speech Recognition: For the English-to-Dravidian pipeline, English speech is transcribed using *Whisper Large-v3* (Radford et al., 2023), while Dravidian-to-Dravidian translation uses *IndicConformer* (Bhogale et al., 2025) for Dravidian ASR.

Text-to-Text Translation: The transcribed text is translated using *IndicTrans2* (Gala et al., 2023), using the English-to-Indic configuration for EN→DL and the Indic-to-Indic configuration for DL→DL translation.

Speech Synthesis: The translated text is synthesized into speech using *Indic Parler-TTS* (Sankar et al., 2025b), conditioned on the translated text and speaker prompts. The TTS component is kept frozen during fine-tuning, as domain-adaptive TTS training requires paired target-language audio with transcriptions at a scale not available in the datasets used here.

End-to-End Architecture: We use *SeamlessM4T-v2* (Seamless Communication Team, 2023) as the End-to-End system, directly mapping source speech to target speech within a unified multilingual architecture without intermediate text generation. The model supports speech-to-speech, speech-to-text, text-to-text, and text-to-speech translation in a single framework, eliminating compounding errors across pipeline stages, though strong performance on low-resource Dravidian pairs remains challenging due to limited parallel speech data.

Parameter-Efficient Fine-Tuning: LoRA (Hu et al., 2022) introduces trainable low-rank matrices into attention layers while keeping original model parameters frozen. LoRA is applied to *Whisper Large-v3* and *IndicConformer* for ASR, to *IndicTrans2* attention layers for machine translation, and to the cross-attention modules (`q_proj`, `k_proj`, `v_proj`, and `out_proj`) of *SeamlessM4T-v2*'s speech encoder and text decoder, optimised with a speech-to-text cross-entropy objective over target text token sequences. After training, LoRA weights are merged with the base model for inference. Across all modules, LoRA rank is fixed at 16 with scaling factor 32 and dropout 0.1, keeping trainable parameters well below 1% of total model size. Both systems are fine-tuned on identical train splits to ensure fair architectural comparison.

The details of the state-of-the-art models used in both architectures are provided in Appendix A.

3 Experimental Setup

We conduct experiments for English-to-Dravidian (EN→DL) and Dravidian-to-Dravidian (DL→DL) speech translation tasks. The inference pipelines are shown in Figure 1.

FLEURS (Conneau et al., 2022) is a multilingual speech benchmark covering 102 languages with roughly 12 hours of aligned speech per language. We use English, Tamil, Telugu, and Kannada subsets with predefined train and test splits. Audio inputs are converted to mono when required and resampled to 16 kHz before inference.

BhasaAnuvaad (Sankar et al., 2025a) contains over 44,000 hours of speech across 14 Indian languages and English. We use the *Mann-ki-Baat* subset with English transcripts and multilingual translations. Since explicit Dravidian-to-Dravidian pairs are unavailable, alignment is performed using LaBSE sentence embeddings, cosine similarity, and Hungarian one-to-one matching (threshold 0.6), yielding approximately 15,284 aligned pairs split 80/10/10 for train/development/test.

Training Parameter Settings Training uses mini-batch optimization with batch size 8, LoRA rank 16, scaling factor 32, dropout 0.1, and learning rate $1e^{-4}$ with warm-up steps. Models are trained for 4000 steps, selecting the best checkpoint based on validation BLEU. Training data sizes are approximately 3,000 utterances per language for FLEURS and 12,200 per language for Mann-ki-Baat (80% of 15,284 aligned pairs), using identical splits for

System	Setting	Lang	BLEU↑		chrF++↑		COMET↑		WER↓		CER↓	
			FL	MKB	FL	MKB	FL	MKB	FL	MKB	FL	MKB
Cascaded	Zero-shot	Tamil	13.75	3.17	52.18	38.44	0.8312	0.7682	0.9587	1.0214	0.6024	0.6438
		Telugu	18.96	7.39	59.42	44.83	0.8724	0.7908	0.8832	0.9634	0.5319	0.6012
		Kannada	14.32	8.70	57.04	45.67	0.8569	0.7994	0.9141	0.9784	0.5693	0.6103
	Fine-tuned	Tamil	14.43	5.08	53.91	42.17	0.8441	0.7934	0.9301	0.9812	0.5847	0.6187
		Telugu	19.18	8.02	60.11	46.59	0.8793	0.8091	0.8614	0.9287	0.5174	0.5784
		Kannada	14.33	7.03	58.76	47.22	0.8651	0.8102	0.8972	0.9441	0.5541	0.5872
E2E	Zero-shot	Tamil	12.10	3.02	49.33	36.91	0.8104	0.7541	1.0234	1.0687	0.6512	0.6741
		Telugu	15.72	6.01	55.87	42.17	0.8391	0.7713	0.9617	1.0118	0.5843	0.6334
		Kannada	12.88	6.24	53.29	43.11	0.8238	0.7762	0.9934	1.0312	0.6187	0.6448
	Fine-tuned	Tamil	13.54	4.11	51.07	39.88	0.8219	0.7716	0.9883	1.0341	0.6215	0.6512
		Telugu	16.84	7.14	57.04	43.98	0.8509	0.7849	0.9248	0.9812	0.5611	0.6141
		Kannada	13.67	7.01	54.88	44.88	0.8349	0.7883	0.9612	0.9987	0.5978	0.6234

Table 1: English→Dravidian speech translation results on FLEURS (FL) and Mann-Ki-Baat (MKB). Cascaded: Whisper Large-v3 → IndicTrans2 → Indic Parler-TTS. E2E: SeamlessM4T-v2.

System	Setting	Pair	BLEU↑		chrF++↑		COMET↑		WER↓		CER↓	
			FL	MKB	FL	MKB	FL	MKB	FL	MKB	FL	MKB
Cascaded	Zero-shot	Ta→Te	8.13	6.05	46.41	41.32	0.7987	0.7562	0.8346	0.8734	0.6043	0.6521
		Ta→Kn	6.73	4.98	46.07	40.91	0.7991	0.7534	0.8569	0.8926	0.6187	0.6674
		Te→Ta	7.38	5.63	45.22	40.74	0.7824	0.7442	0.8216	0.8661	0.5882	0.6359
		Te→Kn	5.81	4.36	44.06	39.72	0.7736	0.7372	0.8472	0.8854	0.5714	0.6271
		Kn→Ta	6.91	5.12	45.64	40.86	0.7916	0.7491	0.8182	0.8578	0.5965	0.6413
		Kn→Te	6.11	4.78	44.71	40.18	0.7832	0.7415	0.8305	0.8729	0.5821	0.6312
	Fine-tuned	Ta→Te	10.12	7.84	48.01	43.65	0.8212	0.7816	0.7821	0.8295	0.5529	0.6104
		Ta→Kn	8.74	6.42	47.26	42.68	0.8161	0.7765	0.8049	0.8489	0.5672	0.6237
		Te→Ta	9.11	6.87	47.08	42.97	0.8073	0.7693	0.7704	0.8187	0.5463	0.5998
		Te→Kn	7.96	5.98	46.44	41.95	0.8011	0.7641	0.7961	0.8397	0.5286	0.5862
		Kn→Ta	9.04	6.73	47.63	43.14	0.8196	0.7776	0.7687	0.8191	0.5487	0.6025
		Kn→Te	8.51	6.21	47.11	42.56	0.8123	0.7728	0.7819	0.8284	0.5361	0.5906
E2E	Zero-shot	Ta→Te	6.84	4.91	44.31	39.61	0.7759	0.7346	0.8527	0.8961	0.6221	0.6715
		Ta→Kn	5.42	3.98	43.64	38.92	0.7682	0.7284	0.8682	0.9083	0.6347	0.6846
		Te→Ta	6.06	4.62	43.77	39.04	0.7704	0.7318	0.8417	0.8873	0.6119	0.6597
		Te→Kn	4.74	3.61	42.68	37.96	0.7621	0.7244	0.8585	0.8997	0.6278	0.6751
		Kn→Ta	5.61	4.21	43.36	38.74	0.7694	0.7297	0.8335	0.8775	0.6044	0.6518
		Kn→Te	4.98	3.84	42.93	38.31	0.7654	0.7261	0.8466	0.8907	0.6166	0.6647
	Fine-tuned	Ta→Te	8.97	6.42	46.51	41.98	0.8036	0.7615	0.8034	0.8467	0.5713	0.6198
		Ta→Kn	7.72	5.62	45.97	41.23	0.7958	0.7551	0.8181	0.8614	0.5868	0.6342
		Te→Ta	8.54	6.03	46.18	41.41	0.8013	0.7584	0.7946	0.8408	0.5617	0.6117
		Te→Kn	7.21	5.23	45.24	40.56	0.7914	0.7516	0.8078	0.8532	0.5791	0.6265
		Kn→Ta	8.16	5.84	45.76	41.02	0.7991	0.7567	0.7858	0.8336	0.5569	0.6046
		Kn→Te	7.52	5.41	45.37	40.78	0.7943	0.7544	0.7979	0.8442	0.5682	0.6153

Table 2: Dravidian→Dravidian speech translation results on FLEURS (FL) and Mann-Ki-Baat (MKB). Ta = Tamil, Te = Telugu, Kn = Kannada.

both Cascaded and E2E systems.

Evaluation Metrics We evaluate translation quality using BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (WMT22 COMET-DA (Rei et al., 2020)). For the End-to-End system, chrF++ and COMET are computed using the SeamlessM4T-v2 S2TT output head, ensuring metric-level parity with Cascaded systems. Speech quality is evaluated via Back-ASR, where generated speech is re-transcribed using IndicConformer and scored by WER and CER; values greater than 1.0 indicate insertion-dominated error ratios, and since IndicConformer is used for both systems, a small bias toward Cascaded models may exist; neutral ASR evaluation is left for future work. To diagnose error propagation, we additionally report stand-alone ASR WER and MT BLEU on gold versus ASR-transcribed input (Table 3), enabling module-level assessment of Cascaded S2S performance.

Dataset preprocessing details are provided in Appendix B.

Direction	Dataset	ASR WER↓		MT BLEU (gold)↑		MT BLEU (ASR)↑	
		Zero-shot	Fine-tuned	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
EN→Tamil	FLEURS	0.124	0.089	17.42	18.11	14.81	15.63
EN→Tamil	MKB	0.198	0.151	5.84	7.23	4.02	5.94
EN→Telugu	FLEURS	0.117	0.082	22.31	22.74	19.87	20.46
EN→Telugu	MKB	0.191	0.143	9.87	10.42	8.14	9.08
EN→Kannada	FLEURS	0.131	0.094	18.07	18.26	15.41	15.82
EN→Kannada	MKB	0.204	0.158	11.23	10.14	9.72	8.29
Ta→Te	FLEURS	0.213	0.178	9.84	11.67	8.42	10.31
Te→Ta	FLEURS	0.221	0.184	9.17	10.83	7.89	9.56
Kn→Ta	FLEURS	0.218	0.181	8.74	10.91	7.41	9.62

Table 3: Cascaded pipeline diagnostics. ASR WER is computed on test-set transcripts before and after LoRA fine-tuning. MT BLEU (gold): gold transcript input; MT BLEU (ASR): ASR hypothesis. The gap between the two MT BLEU columns quantifies error propagation from ASR into the translation step. EN ASR uses Whisper Large-v3; Dravidian ASR uses IndicConformer.

4 Experimental Results and Discussion

Tables 1 and 2 present EN→DL and DL→DL results, while Table 3 reports component-level diagnostics for the Cascaded pipeline.

Effect of Fine-tuning: Fine-tuning improves performance across most language pairs. In EN→DL translation, Tamil BLEU improves from 13.75 to 14.43 on FLEURS (4.9%) and from 3.17 to 5.08 on Mann-ki-Baat (60.3%), while Telugu WER decreases from 0.9634 to 0.9287. In DL→DL translation, Tamil→Telugu BLEU improves from 8.13 to 10.12 (24.5%) and Tamil→Kannada from 6.73 to 8.74 (29.9%), with similar gains across other Dravidian pairs, demonstrating the effectiveness of LoRA fine-tuning.

Cascaded Modular vs. End-to-End: Cascaded systems outperform End-to-End models across most settings. For EN→Telugu on FLEURS, Cascaded zero-shot BLEU (18.96) exceeds End-to-End (15.72) by 20.6%, persisting after fine-tuning (19.18 vs. 16.84). Similarly, Tamil→Telugu exceeds End-to-End both before (8.13 vs. 6.84) and after fine-tuning (10.12 vs. 8.97). However, some margins are small—Kannada on FLEURS after fine-tuning differs by only 0.66 BLEU (14.33 vs. 13.67)—and should be interpreted cautiously without significance testing. These findings align with prior work showing that adapted End-to-End systems can approach cascaded models in higher-resource settings (Sankar et al., 2025a).

Component-Level Diagnostics: ASR fine-tuning reduces WER substantially: English ASR WER on FLEURS improves from 0.124 to 0.089, and Tamil from 0.213 to 0.178. MT BLEU on gold transcripts consistently exceeds ASR-input BLEU, quantifying error propagation; for EN→Tamil on FLEURS, MT BLEU drops from 17.42 (gold) to 14.81 (ASR input), a reduction of 2.61 points.

Speech Quality: Back-ASR evaluation confirms improved speech quality after fine-tuning: for Tamil→Telugu, WER decreases from 0.8346 to 0.7821 and CER from 0.6043 to 0.5529, indicating improved intelligibility, though naturalness and prosody remain unassessed.

Overall, Cascaded Modular pipelines remain strong baselines for low-resource Dravidian S2S translation, while End-to-End systems offer architectural simplicity with increasing competitiveness.

5 Conclusion and Future Work

We presented a comparative study of Cascaded Modular and End-to-End speech-to-speech translation for Tamil, Telugu, and Kannada. Cascaded systems generally outperform End-to-End models across most settings, though margins are small for some pairs and no significance testing was performed. LoRA fine-tuning yields BLEU gains of approximately 20–30% for Dravidian pairs, and component-level diagnostics confirm that ASR fine-tuning meaningfully reduces error propagation into the translation stage, highlighting the effectiveness of modular architectures for low-resource speech translation.

As future work, we aim to extend this research to code-switched speech processing for Dravidian languages by constructing code-switched datasets through merging of monolingual speech segments, investigate zero-shot ASR with parameter-efficient fine-tuning, and incorporate human evaluation of TTS naturalness alongside acoustic quality metrics such as STOI and PESQ.

6 Limitations

First, experiments are restricted to Tamil, Telugu, and Kannada; evaluating additional Dravidian languages would provide broader multilingual insights. Second, GPU memory constraints limited training to a smaller subset of Mann-ki-Baat, which may affect performance compared to full-scale training. Third, FLEURS contains clean read speech while Mann-ki-Baat consists of broadcast-style speech, introducing domain variability that may affect generalization. Fourth, the automatic DL→DL alignment from Mann-ki-Baat has not been human-validated, and noisy pairs remain a potential source of error. Fifth, using IndicConformer for Back-ASR evaluation of both systems may introduce a small bias favouring the Cascaded pipeline; neutral ASR evaluation is left for future work. Sixth, Indic Parler-TTS was not fine-tuned due to insufficient paired audio-transcript data, and the effect of TTS adaptation on overall S2S performance remains an open question. Seventh, no human evaluation of speech naturalness or prosodic quality was conducted; back-ASR captures intelligibility but not speaker naturalness. Finally, statistical significance tests were not conducted; differences below $\approx$1 BLEU point should be interpreted with caution.

References

Kaushal Santosh Bhogale, Deovrat Mehendale, Tahir Javed, Devbrat Anuragi, Sakshi Joshi, Sai Sundaresan, Aparna Ananthanarayanan, Sharmistha Dey, Anusha Srinivasan, Abhigyan Raman, Mitesh M. Khapra, and Anoop Kunchukuttan. 2025. Mahadhwani: Bringing parity in speech recognition for low-resource indian languages. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

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

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 5036–5040, Shanghai, China. International Speech Communication Association (ISCA).

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M. Khapra, and Raj Dabre. 2025a. 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)*, Vienna, Austria. Association for Computational Linguistics.

Ashwin Sankar, Yoach Lacombe, Sherry Thomas, Praveen Srinivasa Varadhan, Sanchit Gandhi, and

Mitesh M. Khapra. 2025b. Rasmalai: Resources for adaptive speech modeling in indian languages with accents and intonations. In *Proceedings of the 26th Annual Conference of the International Speech Communication Association (INTERSPEECH 2025)*, Rotterdam, Netherlands. International Speech Communication Association (ISCA). AI4Bharat, IIT Madras and HuggingFace.

Seamless Communication Team. 2023. Seamlessm4t: Massively multilingual and multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Appendix Section

A Details of the Key Models

Whisper (Radford et al., 2023): A large multilingual automatic speech recognition (ASR) model trained on approximately 680,000 hours of weakly supervised speech data collected from the internet. In our analysis, we use *Whisper Large-v3* to transcribe English speech in the English-to-Dravidian translation pipeline.

IndicConformer (Bhogale et al., 2025): A multilingual ASR model designed for Indian languages. It is based on the Conformer architecture (Gulati et al., 2020), which combines convolutional neural networks with transformer encoders to effectively capture both local and global speech patterns. IndicConformer supports transcription across the 22 scheduled Indian languages. In this work, we use *IndicConformer* for speech recognition in the Dravidian-to-Dravidian translation pipeline and for Back-ASR evaluation of both systems.

IndicTrans2 (Gala et al., 2023): A transformer-based multilingual neural machine translation system that supports translation across Indic languages. The model leverages script unification and multilingual training to improve translation quality for low-resource languages. In cascaded speech translation pipelines, *IndicTrans2* is used to translate the recognized text into the target Dravidian language.

Indic Parler-TTS (Sankar et al., 2025b): A multilingual text-to-speech system designed for Indian languages. The model generates natural and expressive speech conditioned on textual prompts and supports several Indic languages. In our work, *Indic Parler-TTS* is used in the final stage of the cascaded pipeline to synthesize speech in the target Dravidian language from the translated text.

SeamlessM4T (Seamless Communication Team, 2023): A massively multilingual and multimodal translation model that performs speech-to-speech, speech-to-text, text-to-text, and text-to-speech translation within a single unified architecture. In this work, we use **SeamlessM4T-v2** as an end-to-end baseline for speech-to-speech translation.

B Details of Experimental Setup

Datasets Preprocessing We conduct experiments using two multilingual speech datasets: **FLEURS** (Conneau et al., 2022) and the **Mann-ki-Baat**, a subset of the BhasaAnuvaad collection[2] (Sankar et al., 2025a). Since the datasets differ in structure and annotation format, different preprocessing strategies are applied before training and evaluation.

FLEURS (Conneau et al., 2022) is a multilingual speech dataset containing speech recordings and aligned transcriptions across more than 100 languages. We use the English and Dravidian language subsets (Tamil, Telugu, and Kannada). Since the dataset is n-way parallel, samples across languages share the same identifier; we align English and target language sentences using their unique IDs and remove samples with missing or empty transcriptions. The predefined training and test splits are used directly: training for fine-tuning and test for both zero-shot and fine-tuned evaluation. Audio inputs are decoded, converted to mono when required, and resampled to 16 kHz before ASR processing.

The Mann-ki-Baat dataset contains speech recordings from the Mann-ki-Baat radio program with English transcripts and translations in multiple Indian languages, but does not provide explicit Dravidian-to-Dravidian pairs. We construct parallel data via automatic alignment using `LaBSE` sentence embeddings (`sentence-transformers/LaBSE`) (Reimers and Gurevych, 2019). Cosine similarity scores are computed between sentence pairs, followed by Hungarian one-to-one matching. Only pairs exceeding a cosine similarity threshold of 0.6 are retained, yielding approximately 15,284 aligned pairs and enabling the construction of semantically consistent pseudo-parallel Dravidian language pairs. The dataset is shuffled with a fixed random seed (42) and split into train/development/test sets using an 80/10/10 ratio, with identical splits reused across both systems for fair comparison. For English-to-Dravidian translation, source audio is directly used for ASR transcription, while samples with missing audio, empty transcripts, or target translations are discarded before splitting. Audio segments are decoded, converted to mono if necessary, and resampled to 16 kHz to ensure consistent input format across both systems, as Whisper, IndicConformer, and SeamlessM4T-v2 all require fixed-rate mono audio for stable training and inference.

[2] https://huggingface.co/datasets/ai4bharat/Mann-ki-Baat