IEEE BIG DATA 2024
December 15-18 • Washington DC, USA

IEEE International Conference on Big Data 2024

BanglaDialecto: An End-to-End AI-Powered Regional Speech Standardization

Md. Nazmus Sadat Samin*, Jawad Ibn Ahad*, Tanjila Ahmed Medha, Fuad Rahman, Mohammad Ruhul Amin, Nabeel Mohammed, Shafin Rahman

Outline

- Introduction
  - Background
  - Challenges in Dialect Standardization
  - Existing Limitations
- Methodology Overview
- Experimental Results
- Conclusion and Future Work

Background

The focus of the Study

- Recognizing Bangladeshi dialects.
- Converting diverse Bengali accents to standardized formal Bangla speech.

Background

The focus of the Study

- Recognizing Bangladeshi dialects.
- Converting diverse Bengali accents to standardized formal Bangla speech.

Key Concepts

- Dialects are defined by phonetics, pronunciation, and lexicon.

Background

The focus of the Study

- Recognizing Bangladeshi dialects.
- Converting diverse Bengali accents to standardized formal Bangla speech.

Key Concepts

- Dialects are defined by phonetics, pronunciation, and lexicon.

Significance

- Bangla: 5th most spoken language globally (Approximately 160M speakers).
- Around 55 distinct dialects.

Challenges in Dialect Standardization

Dialectal Variations

- Pronunciation, phonetics, and vocabulary differences due to geography and socioeconomic factors.

Data Scarcity

- Lack of large-scale, diverse dialect speech datasets.

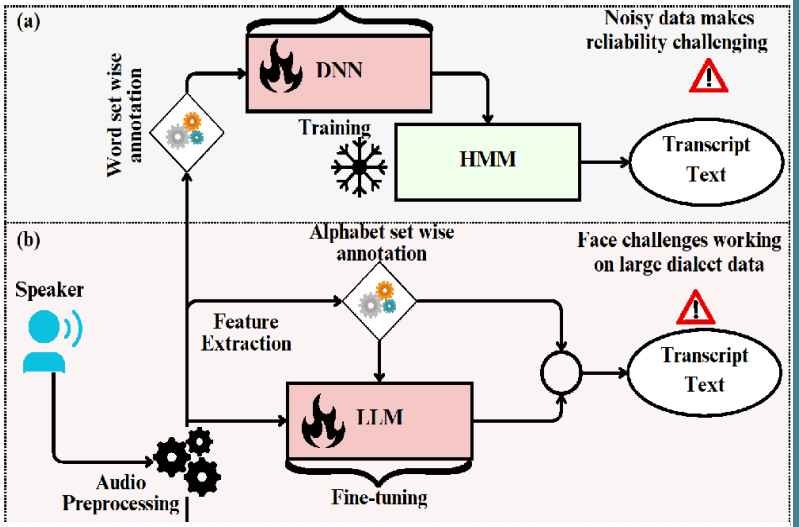
Technical Complexity

- Managing large dialect data for fine-tuning.
- Building accurate ASR (Automatic Speech Recognition) and Machine Translation models for dialectal speech.

Existing Limitations

Traditional Approaches

- DNN-based ASR relies heavily on conventional methods, often with limited success in dialect recognition.



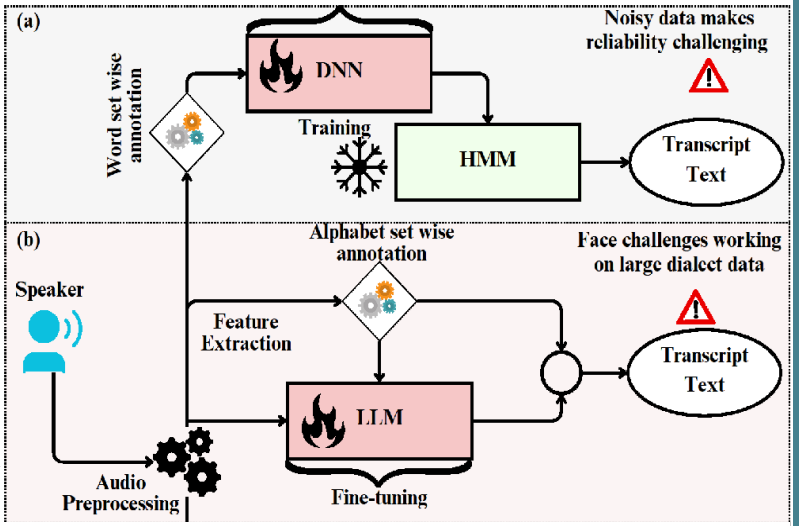
Existing Limitations

Traditional Approaches

- DNN-based ASR relies heavily on conventional methods, often with limited success in dialect recognition.

Limited Integration

- Previous efforts focused on ASR, MT, or TTS tasks individually.
- Few studies integrate these into an end-to-end pipeline for dialect standardization.



Existing Limitations

Traditional Approaches

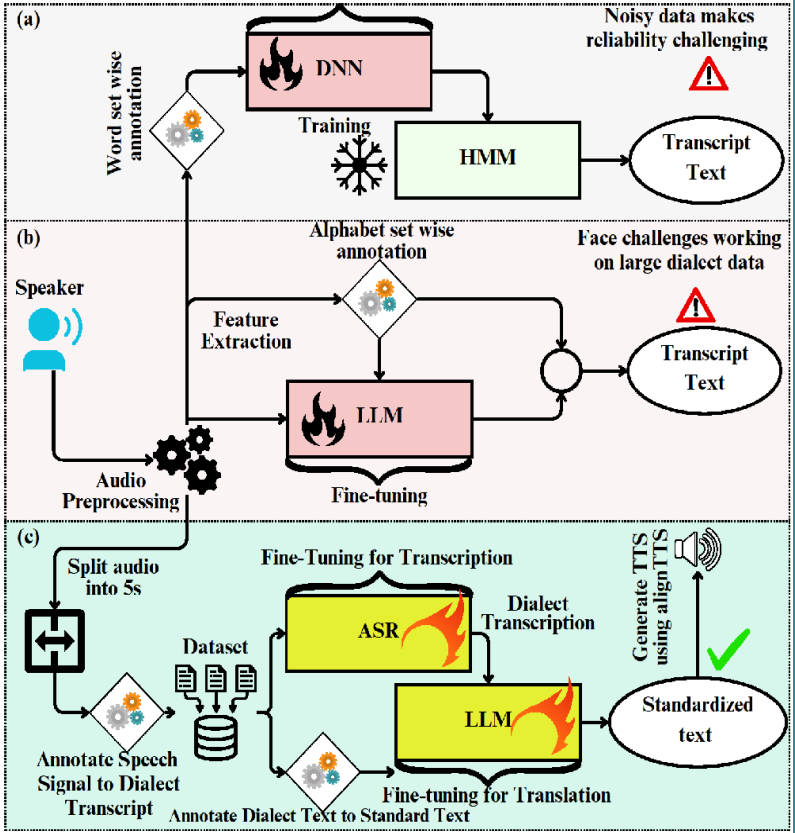
- DNN-based ASR relies heavily on conventional methods, often with limited success in dialect recognition.

Limited Integration

- Previous efforts focused on ASR, MT, or TTS tasks individually.
- Few studies integrate these into an end-to-end pipeline for dialect standardization.

Model Performance

- Pretrained models often fail to capture dialect-specific linguistic nuances without extensive fine-tuning.



Outline

- Introduction
- Methodology Overview
  - Noakhali Dialect Dataset (NDD)
  - Preprocessing
  - BanglaDialecto System
- Experimental Results
- Conclusion and Future Work

Noakhali Dialect Dataset (NDD)

Overview

- First Dataset for the Noakhali dialect.
- Collected from 24 speakers (18 males, 6 females, aged 18-28).

Dataset Details

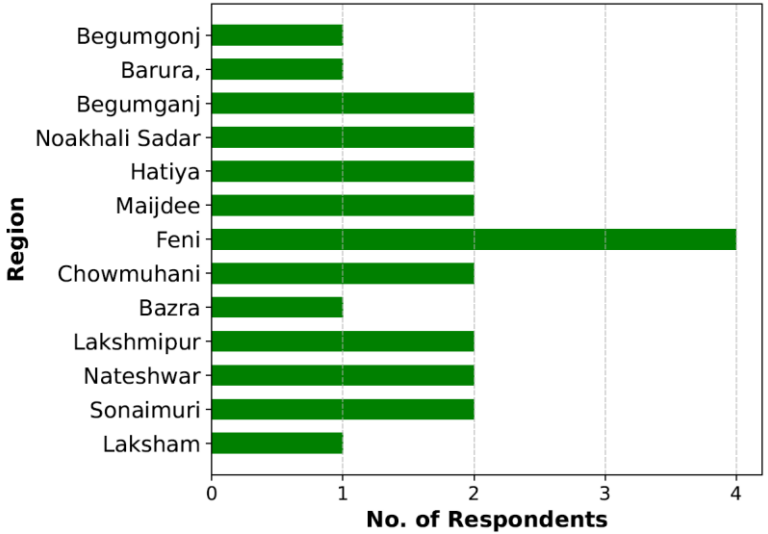
- **Speech Data:** 10 hours of audio.

Annotations

- Noakhali dialect text.
- Standard Bangla text.

Sources

- YouTube videos, Facebook groups, and interviews



Preprocessing

Step 1: Denoising

- Noise removed using *noisereduce* library for clarity.

Step 2: Segmentation

- Speech data segmented into 5-second chunks.
- Aligned text annotations split accordingly for synchronization.

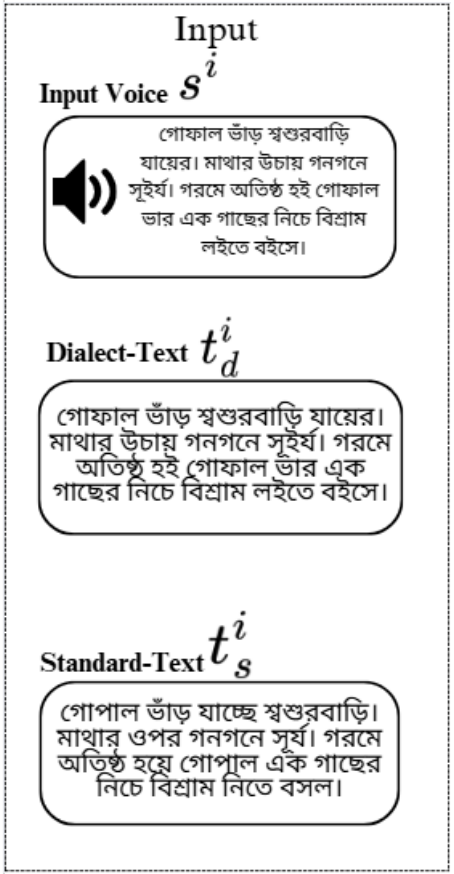
$$N^i = \left[\frac{T^i}{5}\right]$$

Step 3: Text Splitting

- Dialect and standard texts segmented for fine-tuning models on aligned speech chunks.

$$t^i{}_d = \{t^i{}_{d1}, t^i{}_{d2}, \ldots\ldots\ldots, t^i{}_{dN}\}$$
$$t^i{}_s = \{t^i{}_{s1}, t^i{}_{s2}, \ldots\ldots\ldots, t^i{}_{sN}\}$$

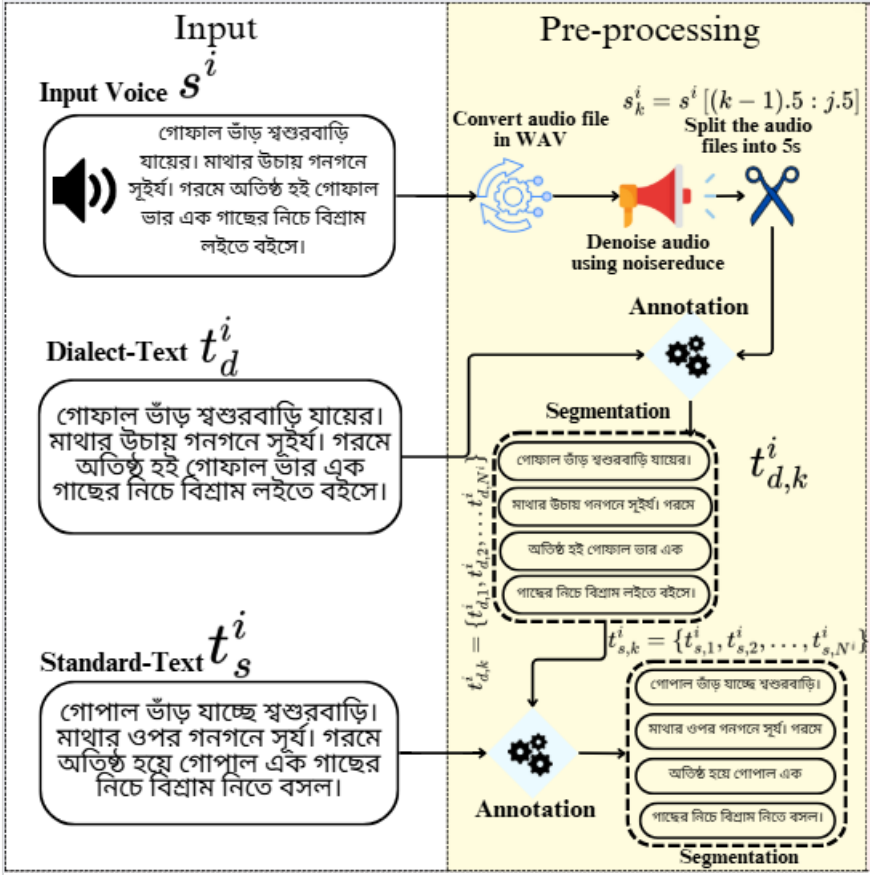
BanglaDialecto System



Input dialect speech signals S^i are converted into wav form, then it undergoes the process of noise reduction and splitting into manageable 5-second speech segments.

Each S^i contains it's corresponding dialect text t_d^i and standard Bangla text t_s^i.

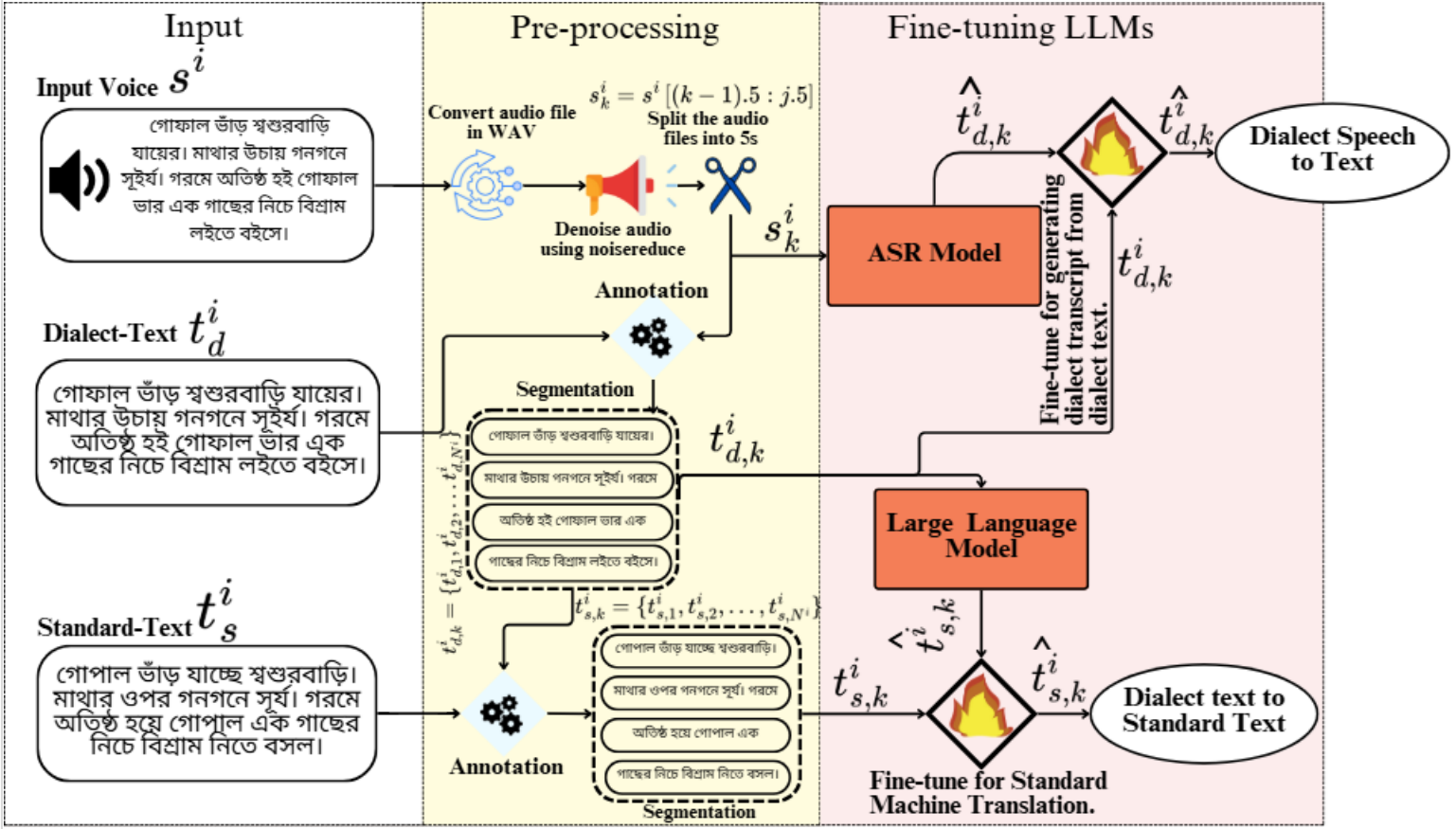
BanglaDialecto System



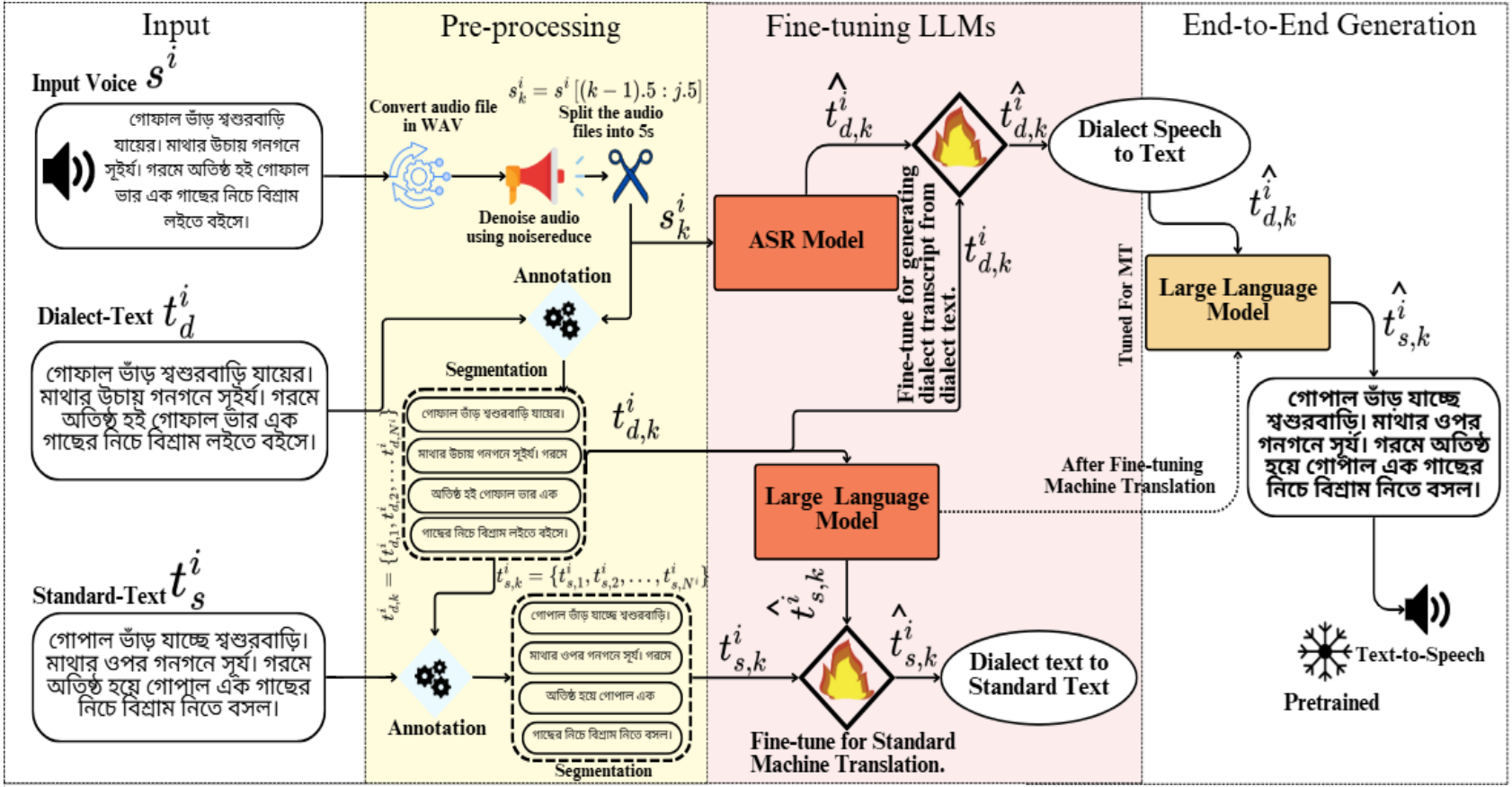
After Pre-processing, each S^i segmented in 5-second, s_k^i, dialect text t_d^i and standard text t_s^i then segmented in corresponding chunks $t_{d,k}^i$ and $t_{s,k}^i$.

The segment s_k^i and $t_{d,k}^i$ are used to fine-tune the ASR to predict and transcript dialect speech s_k^i into dialect text t_d^i .The other segment $t_{s,k}^i$ alongside with $t_{d,k}^i$ is used to fine-tune the LLMs for MT from dialect text to standard Bangla text.

BanglaDialecto System



BanglaDialecto System



BanglaDialecto System

Speech-to-Text (ASR)

- Fine-tuned Whisper models (base to large V2) for dialect transcription.
- Input: Dialect speech. Output: Dialect text.

Text-to-Text Translation (LLM)

- Models: BanglaT5, mBART50, IndicBART, mT5.
- Fine-tuned for translating dialect text into standard Bangla text.
- Input: Dialect text. Output: Standard Bangla text.

End-to-End Pipeline

- Whisper ASR → Dialect Text → BanglaT5 → Standard Text → AlignTTS for speech synthesis.

Outline

- Introduction
- Methodology Overview
- Experimental Results
  - Setup
  - Results and Analysis
  - Discussion
- Conclusion and Future Work

Setup

➢ Dataset Details

- **Noakhali Dialect Dataset (NDD):** 10 hours of speech, 7,200 samples.
- **Split:** 6270 (training), 810 (validation), 120 (testing).

➢ Fine-Tuning

- ASR: Whisper models (base to large V2).
- Translation: BanglaT5, mBART50, IndicBART, mT5.

➢ Hardware & Framework

- **GPU:** NVIDIA Tesla A100.
- **Framework:** PyTorch.

Results and Analysis

Task	Models	Parameter	Pre Trained version			Fine-tuned version		
			CER(%)↓	WER(%)↓	BLUE Score(%)↑	CER(%)↓	WER(%)↓	BLUE Score(%)↑
Dialect Speech-to-text (ASR)	Sequential Transformer	24M	87	179.8	-			
	Whisper-base	74M	230.8	232.2	-	20.6	47.1	-
	Whisper-small	244M	389.8	411.9	-	2.0	4.0	-
	Whisper-medium	769M	259.2	169.0	-	1.5	2.0	-
	Whisper-large V2	1550M	135.2	167.5	-	**0.8**	**1.5**	-
Dialect text Translation (MT)	mBART50	611M	248.5	612.8	0.45	79.8	416.8	3.0
	IndicBART	244M	166.5	409.8	11.9	22.8	118.6	27.7
	mT5(base)	582M	275.4	405.4	5.6	34.4	59.2	18.6
	BanglaT5	247M	178.9	281.6	22.7	**21.3**	**38.2**	**41.6**

ASR Performance

- Best model: Whisper-large V2.
- CER: 0.8%.
- WER: 1.5%.

Translation Performance

- Best model: BanglaT5.
- BLEU Score: 41.6.

Key Insights: Fine-tuning significantly improves performance for both ASR and MT tasks.

Results and Analysis

ASR				MT				
Methods	**Language**	**CER(%)↓**	**WER(%)↓**	**Methods**	**Language**	**CER(%)↓**	**WER(%)↓**	**BLEU(%)↑**
Messaoudi el al. [1]	Tunisian dialect	18.7	24.4	Faria et al. [8]	Mymensingh dialect	8.23	15.48	**69.06**
Ying et al. [2]	Sichuan dialect	25.24	57.02	Faria et al. [9]	Noakhali dialect	20.3	38.7	47.43
Alam el al [3]	Standard Bangla	13.856	39.29	Kchaou et al. [10]	Tunisian dialect	-	-	60.00
Kabir et al. [4]	Sylhet Dialect	-	14.89	Hamed et al. [11]	Arabic	-	-	18
Nasr et al. [5]	Arabian dialect	20	30	Baruah et al. [12]	Assamese	-	-	50.19
Phung et al. [6]	Vietnamese	4.79	2.79	Adiputra et al. [13]	Japanese-Indonesian	-	-	8.78
Safieh et al. [7]	Jordanian dialect	26.4	51.50	Prama et al. [14]	Sylhet Dialect	-	-	57.4
Ours	Noakhali Dialect	**0.8**	**1.5**	Ours	Noakhali dialect	**20.2**	**38.2**	41.6

Comparative evaluation of pre-trained and fine-tuned models on speech-to-text and dialect text-to-standard text translation tasks. The models are assessed using character error rate (CER), word error rate (WER), and BLEU score. For ASR, Whisper-large V2 shows the best results with a CER of 0.8% and a WER of 1.5%. For DTT, BanglaT5 outperforms others with a BLEU score of 41.6, demonstrating superior performance in translation tasks. Fine-tuning consistently improves model accuracy across all metrics.

[1] A. Messaoudi, H. Haddad, C. Fourati, M. B. Hmida, A. B. E. Mabrouk, and M. Graiet, "Tunisian dialectal end-to-end speech recognition based on deepspeech,"
[2] W. Ying, L. Zhang, and H. Deng, "Sichuan dialect speech recognition with deep lstm network,"
[3] S. Alam, A. Sushmit, Z. Abdullah, S. Nakkhatra, M. Ansary, S. M. Hossen, S. M. Mehnaz, T. Reasat, and A. I. Humayun, "Bengali common voice speech dataset for automatic speech recognition,"
[4] S. Kabir, N. Nahar, S. Saha, and M. Rashid, "Automatic speech recognition for biomedical data in bengali language,"
[5] S. Nasr, R. Duwairi, and M. Quwaider, "End-to-end speech recognition for arabic dialects,"
[6] A. A. Safieh, I. A. Alhaol, and R. Ghnemat, "End-to-end jordanian dialect speech-to-text self-supervised learning framework,"
[7] N. H. Phung, D. T. Dang, K. D. Ta, K. T. A. Nguyen, T. K. Tran, and C. T. Nguyen, "Enhancing whisper model for vietnamese specific domain with data blending and lora fine-tuning,"
[8] F. T. J. Faria, M. B. Moin, A. A. Wase, M. Ahmmed, M. R. Sani, and T. Muhammad, "Vashantor: a large-scale multilingual benchmark dataset for automated translation of bangla regional dialects to bangla language,"
[9] S. Kchaou, R. Boujelbane, and L. Hadrich, "Hybrid pipeline for building arabic tunisian dialect-standard arabic neural machine translation model from scratch,"
[10] H. Hamed, A. M. Helmy, and A. Mohammed, "Deep learning approach for translating arabic holy quran into italian language,"
[11] M. A. Sulaeman and A. Purwarianti, "Development of indonesian japanese statistical machine translation using lemma translation and additional post-process,"
[12] T. T. Prama and M. M. Anwar, "Sylheti to standard bangla neural machine translation: A deep learning-based dialect conversion approach."

Results and Analysis

End-to-end generation from our fine-tuned models. For each input dialect voice, the model Whisper-large V2 generated the dialect texts, listed in the ASR column, from which the BanglaT5 model generated standard Bangla Text. The second table is given as a reference for English pronunciation and Translation.

Input Voice	ASR (Whisper Large V2)	MT (BanglaT5)
আন্নে কোনাই যান	আন্নে কোনাই যান	আপনি কোথায় যাচ্ছেন
সাইন্সের তুন ভালা রেজাল্ট কইচ্ছি	সাইন্সের তুন ভালা রেজাল্ট কইচ্ছি	সায়েন্স থেকে ভালো রেজাল্ট করেছি
বাইত কিছু কাম আছে হেয়াল্লাই আইছি।	বাইত কিছু কাম আছে হেয়াল্লাই আইছি।	বাড়িতে কিছু কাজ আছে তাই এসেছি
একমাসের মত যাইতাম ই হারিনা	একমাসের মত যাইতাম ই হারিনা	একমাসের মত যেতেই পারিনি
হেতে কত্তন ধরি বই আছে	হেতে কত্তন ধরি বই আছে	সে কতক্ষন ধরে বসে আছে

Input Voice	ASR (Whisper Large V2)	MT (BanglaT5)	English (Translation)
Anne konai jan	Anne konai jan	Apne kothay jan	Where are you going
Science tun vala result koichi	Science tun vala result koichi	Science theke valo result korechi	Have done a good result from Science
Bait kichu kam ase hellai aichi	Bait kichu kam ase hellai aichi	Barite kichu kaj ase tai esechi	Have some work at home so I came
Ekmasher moto jaitami harina	Ekmasher moto jaitami harina	Ekmasher moto jatei parina	Can't go since one month
Hete koto khn dhori boi ase	Hete koto khn dhori boi ase	She koto kokhn dhore bose ache	How long has he been sitting

Discussion

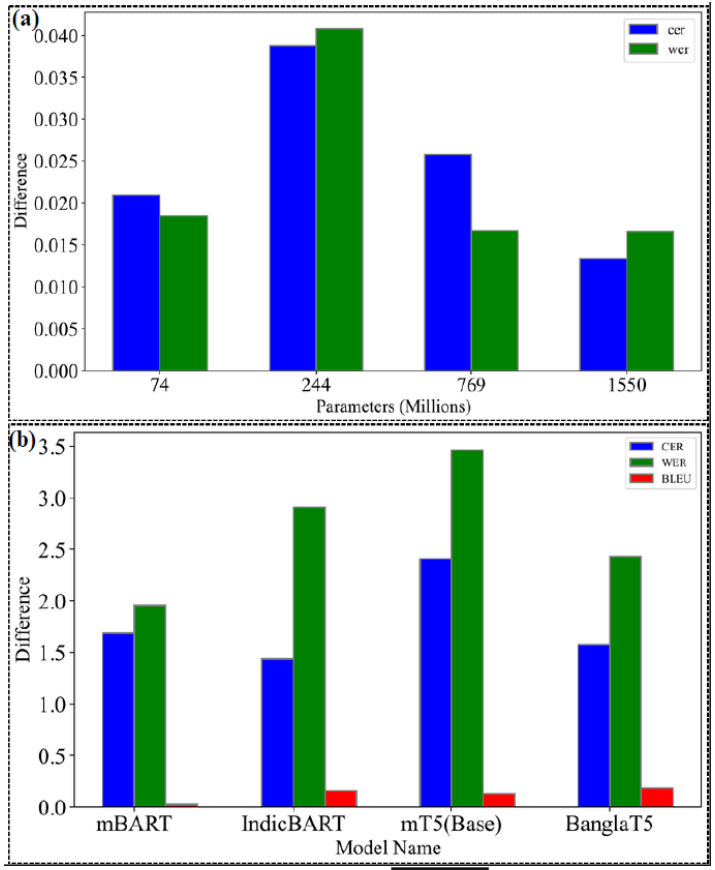
Ablation Study

➢ ASR Models

- Performance improves with model size.
- Best: **Whisper-large V2** (CER: 0.8%, WER: 1.5%).

➢ Translation Models

- Best: **BanglaT5** (BLEU: 41.6).
- Other models like mBART50 and IndicBART show lower performance.



Discussion

➢ Key Contributions

- End-to-end pipeline integrating ASR, MT, and TTS.
- Custom dataset (NDD) for Noakhali dialect.

➢ Limitations

- Focus on a single dialect.
- The dataset lacks coverage of other Bangla dialects.

Outline

- Introduction
- Methodology Overview
- Experimental Results
- Conclusion and Future Work

Conclusion

Summary

- Developed an end-to-end pipeline for Bangla dialect standardization.
- Integrated ASR (Whisper), Machine Translation (BanglaT5), and TTS (AlignTTS).

Impact

- Addresses challenges of dialect variation in Bangla.
- Potential to improve communication in education, media, and other formal contexts.

Future Work

Dataset Expansion

- Include more Bangla dialects to ensure broader applicability.

Multilingual Capabilities

- Extend the system for other low-resource languages.

Enhanced Contextual Understanding

- Improve model performance by capturing nuanced dialectal variations.

Application Development

- Deploy solutions in real-world scenarios like education, healthcare, and public services.

Thank You