Are ASR Foundation Models Generalized Enough to Capture Regional Dialects in Low-Resource Languages?

Tawsif Tashwar Dipto[1], Azmol Hossain[3], Rubayet Sabbir Faruque[2], Md. Rezuwan Hassan[2]
Kanij Fatema[2], Tanmoy Shome[2], Ruwad Naswan[3], Md. Foriduzzaman Zihad[3], Mohaymen Ul Anam[1]
Nazia Tasnim[3], Hasan Mahmud[1], Md. Kamrul Hasan[1], Md. Mehedi Hasan Shawon[2]
Farig Sadeque[2], Tahsin Reasat[3]

[1]Islamic University of Technology [2]Brac University [3]Bengali.AI

Bengali Dialectal ASR: Why It Matters

- Bengali has 270+ million speakers worldwide - 3rd most spoken native language in the world
- Yet current ASR systems (e.g., Whisper) are trained almost exclusively on Standard Colloquial Bengali (SCB)
- Real-life Bengali speech is highly dialectal which lives in a separate feature space than SCB
- Neglecting the dialectal variations in phonology, morphology, syntax and prosody characteristic of authentic speech results in a severe performance degradation for foundation models

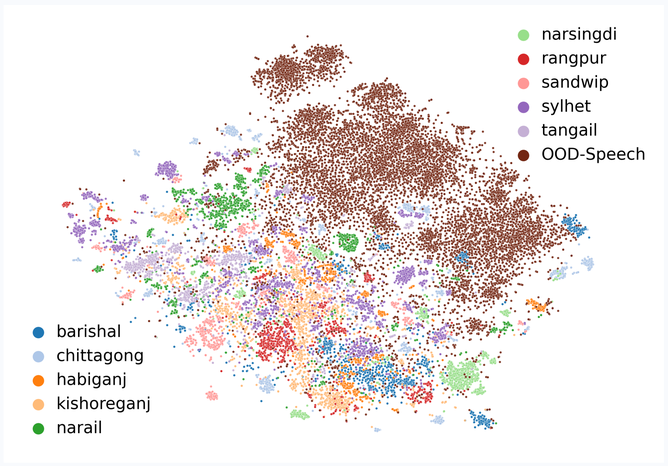

Figure 1: t-SNE of GeMAPS acoustic features shows a clear distribution shift between Standard SCB (brown cluster) and Regional Dialects.

Linguistic Diversity

Our analysis reveals why standard models fail. Bengali dialects are not just "accents" They diverge in every linguistic dimension. We observe massive shifts from the standard dialect:

- Phonology: Aspirated shifts (e.g., Sylheti ক → খ) and heavy nasalization in Chittagong/Sandwip.
- Morphology: Unique clitics (Singular: টা→ডা) and verb conjugations (খেতে পারে → খাইত্তারে).
- Syntax: Subject-Object-Verb deviations. Negation placement shifts (e.g., Chittagonian: যাইনা → ন যাই).
- Lexicon: High Out-of-Vocabulary (OOV) rates. Up to 62% of words in Sandwip are unique compared to SCB.

Region	Regional Sentence	Standard Bengali	English Translation
Rangpur	মোবাইলোত একনা রেকোড করমো। [mobɐ̃ilot ɜknɐ rekod kɔrmo]	মোবাইলে এখন রেকর্ড করবো। [mobɐ̃ile ɜkʰon rekɔrd korbo]	I will record on mobile now.
Kishoreganj	এনো কি লিকছে দ্যাহো। [eno kɪ lɪkcʰe d̪ɜho]	এখানে কি লিখেছে দেখো। [ekʰɐne kɪ lɪkʰecʰe d̪ɜkʰo]	See what is written here.
Narail	তালি তো বাইড়ে গেছে। [t̪elɪ to bɐ̃iɽe gecʰe]	তাহলে তো বেড়ে গেছে। [t̪ɐhole t̪o beɽe gecʰe]	Then it has increased.
Chittagong	কিসু অইতো ন। [kɪcʰu ɔ̃ito nɔ]	কিছু হবে না। [kɪcʰu hɔbe nɐ]	Nothing will happen.
Sandwip	হেতে আইতো ন [heʦe ɐ̃ito nɔ]	সে আসবে না [ʃe ɐʃbe nɐ]	He won't come.
Sylhet	আমরার লাগিও দোয়া খরিও। [ɐmrɐr lɐgiɔ̃ d̪oʊ kʰorɪɔ̃]	আমার জন্যও দোয়া কোরো। [ɐmɐr ɟonnɔ̃ d̪oʊ koro]	Pray for me too.
Habiganj	ইস্কুলের কিতা খইতাম? [ɪskuler kɪt̪ɐ kʰɔ̃it̪ɐm?]	স্কুলে কী বলব? [skule kɪ bolbo?]	What to say at school?
Narsingdi	উল্টা-ফাল্টা কতা ক। [ult̪ɐ-pʰɐlt̪ɐ kɔt̪ɐ kɔ]	উলটা-পাল্টা কথা বলে। [ult̪ɐ-palt̪ɐ kɔt̪ʰɐ bole]	Talk back and forth.
Tangail	এই ছবিডা দ্যাহা অয় নাই। [eɪ̯ cʰobɪd̪ɐ d̪ɜha ɔʲ nɑɪ̯]	এই ছবিটা দেখা হয়নি। [eɪ̯ cʰobɪt̪ɐ d̪ɜkʰɐ hɔʲnɪ]	This image has not been viewed.
Barishal	এহন বাড়ি গ্যালে রাগ অয়। [ehon bɐɽɪ gɜle rɐg ɔʲ]	এখন বাড়ি গেলে রাগ হয়। [ɜkʰon bɐɽɪ gele rɐg hɔʲ]	Now I get angry when I go home.

Table 1: Regional dialectal variations in Bengali.

Contributions

- Ben-10 Dataset: 78 hours, 16,690 clips, 394 speakers across 10 districts.
- High Diversity: Average OOV rate of 59% compared to SCB.
- Linguistic Analysis: Comprehensive mapping of phonological and prosodic shifts.
- Benchmarking: Evaluation of 7 modern ASR models including Whisper, Conformer, and Wav2Vec2.

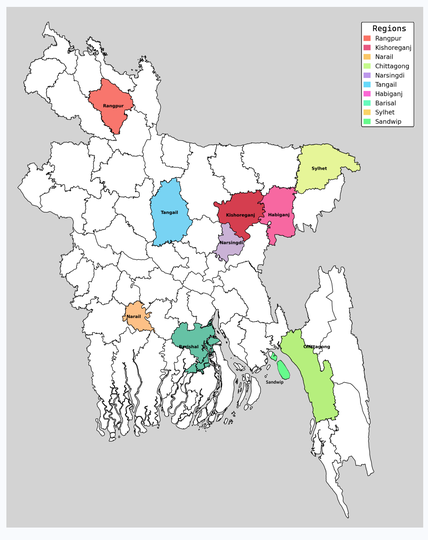

Figure 2: Mapped dialect regions with reference points

Dataset Overview

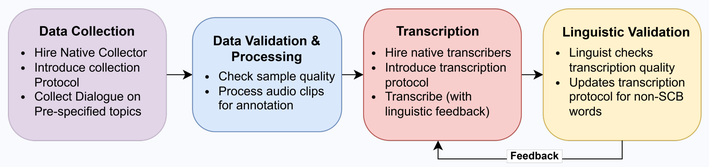

Figure 3: Dataset Creation Workflow: Collection → Validation → Transcription → Linguistic Review.

Regions: Rangpur, Kishoreganj, Narail, Chittagong, Narsingdi, Tangail, Sylhet, Habiganj, Barishal, Sandwip.

- Volume: 78 hours of spontaneous speech.
- Content: 155 conversation topics (family, politics, sports).
- Vocabulary: 62,762 unique words.

Districts	Sample Counts			Duration [H:M]			OOV%			WPM	Contributor
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test		(M/F/B)
Rangpur	1,037	131	130	4:48	0:36	0:35	47.86	37.64	37.99	134.38	27 (19/8/0)
Kishoreganj	1,638	204	206	7:42	0:55	0:58	62.33	53.57	47.14	117.96	47 (25/18/4)
Narail	1,488	183	188	6:52	0:52	0:51	54.74	41.29	43.42	136.81	37 (21/12/4)
Chittagong	1,406	174	177	6:35	0:48	0:47	61.13	56.28	63.33	134.42	41 (15/22/4)
Narsingdi	1,098	136	137	5:04	0:38	0:37	52.24	39.18	37.59	148.53	26 (9/16/1)
Tangail	987	131	132	4:54	0:36	0:35	43.04	24.72	23.06	141.67	36 (18/11/7)
Habiganj	940	117	113	4:20	0:32	0:33	56.55	57.96	54.28	123.47	34 (19/15/0)
Barishal	796	105	105	3:45	0:30	0:30	48.29	43.62	44.86	123.79	26 (6/7/13)
Sylhet	2,903	356	362	13:34	1:50	1:41	63.24	50.36	50.27	126.5	94 (62/30/2)
Sandwip	1,049	129	132	4:48	0:36	0:37	61.91	51.61	52.77	144.12	26 (15/9/2)
Total	13,342	1,666	1,682	6:27	7:58	7:50	72.54	59.72	58.46	131.38	394 (209/148/37)

Table 2: Ben-10 dataset statistics. OOV → words unique to the district that are Out Of Vocabulary in comparison to SCB. M → Male, F → Female, B → Both.

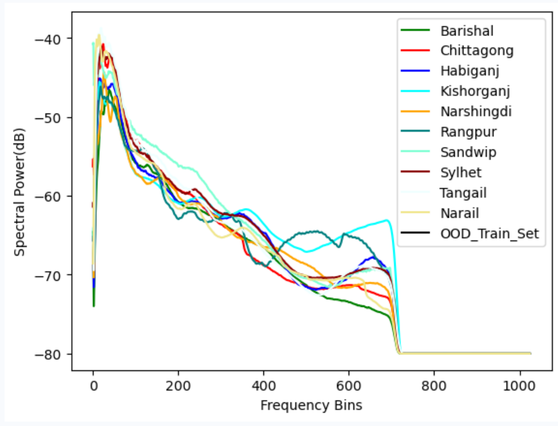

Figure 4: Long-term Spectral Average of Record- ings from different dialects

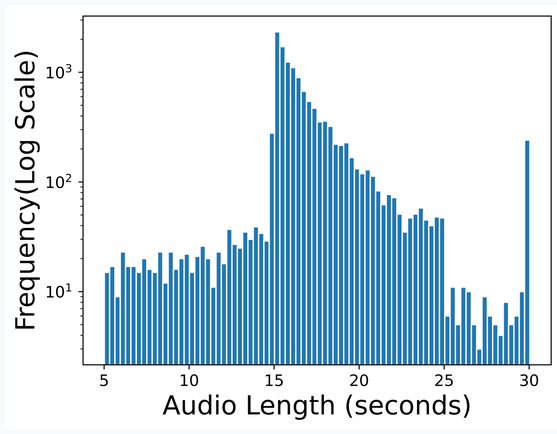

Figure 5: Recording length distribution in Training data

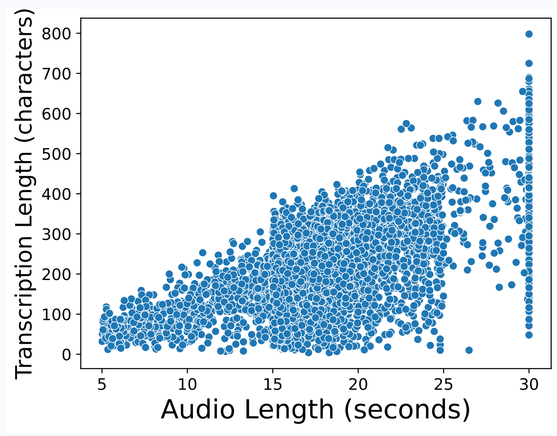

Figure 6: Dialect-wise sentence text length distribution

Methodology

Evaluated Models:

- Foundational: Whisper Large V3, Google Conformer ASR, Hishab Conformer.
- Fine-Tuned: Wav2vec2 (SCB vs Ben-10), Tugstugi (Whisper fine-tuned).

Metrics: WER (Word Error Rate) & CER (Character Error Rate)

Benchmark Results

- Fine-tuning on Ben-10 significantly outperformed standard pre-trained models, achieving substantial improvements in dialectal ASR accuracy across all regions
- Zero-shot performance remains poor, highlighting the critical need for dialect-specific training data

Model	Avg WER ↓	Avg CER ↓
Whisper-Large-V3 (Zero-Shot)	1.13	0.81
Google ASR (API)	1.03	1.03
Hishab Conformer	0.87	0.56
Wav2vec2 (SCB Trained)	0.89	0.62
Tugstugi Whisper (SCB)	0.81	0.52
Wav2vec2 (Ben-10 Finetuned)	0.83	0.53
Tugstugi (Ben-10 Finetuned)	0.70	0.36

Table 3: Fine-tuning on Ben-10 reduces WER by ∼14% compared to the best SCB model.

Regional Variance:

- Best Performance: Sylhet & Tangail (Stronger training data/closer prosody).
- Worst Performance: Chittagong & Sandwip (High OOV and rapid speech).

Why Models Fail: Error Analysis

Foundation models exhibit a strong SCB-Bias. When they encounter a dialect word, they often hallucinate the phonetically closest Standard Bengali word instead of transcribing the actual dialect.

Region	Ground Truth	Tugstugi	Tugstugi (Ben-10)
Rangpur	ধইরছং [d̪o̯ɪrcʰoŋ] গেসনু [gesnu]	দুঃসা [d̪uʃʃʉ] <>	দিতো [d̪ɪt̪o] গেছেনু [gecʰenu]
Kishoreganj	কইতাম [ko̯ɪt̪em] দুইন্নার [d̪u̯ɪnnɐr]	<> দুনিয়ায় [d̪unɪʉʲ]	করোনি [kɔronɪ] দুইন্যের [du̯ɪnner]
Narail	মাইসোতো [mɐ̯ɪʃoʈo] ম্যালাডিক [mɛlɐdɪk]	মায়ের সেতু [mɐer ʃeʈu] মেলাডি [mɛlɐdɪ]	মাইনষে ও তো [mɐ̯ɪnʃe o t̪o] মেলাডি [mɛlɐdɪ]
Chittagong	বেককুনে [bekkune] হারাপ [hɐrɐp]	এখনো [ɛkʰono] <>	একখোনো [ɛkkʰono] খালা [kʰɐlɐ]
Sandwip	লাড়িয়ালান [lɐɪɾɐlɐn] এগগেরে [eggere]	রিলান্দার [rɪlɐnd̪ɐr] <>	হেলার [helɐr] এগেরে [egere]
Sylhet	ছোখো [cʰokʰo] তাইনের [t̪ɐ̯ɪner]	সৌকে [ʃo̯uke] তাই [t̪ɐ̯ɪ]	ছো ওকে [cʰ oke] তাই [t̪ɐ̯ɪ]
Habiganj	আছিলায় [acʰɪlaʲ] খিছুদিন [kʰɪcʰud̪ɪn]	চাইছিলা [cʊ̯ɪcʰɪlɐ] পৃথিবীর [prɪt̪ʰɪbɪr]	চাইছিলায় [cʊ̯ɪcʰɪlɐʲ] ফিতুদিন [fɪt̪ud̪ɪn]
Narsingdi	খেইলাই [kʰɛ̯ɪlɐ̯ɪ] ডাহাত [d̪ɐhɐt̪]	ফেলে [fele] ডাকতেছে [d̪ɐkt̪ecʰe]	ফেলায় [fɛlɐʲ] ডাত [d̪ɐt̪]
Tangail	দ্যাহে [d̪ɐhe] থোয় [t̪ʰoʲ]	দেহে [dehe] হয় [hoʲ]	দেহে [dehe] দেখায় [d̪ekʰɐʲ]
Barishal	কাইলগো [kɐ̯ɪlgo] বাইরাইয়া [bɐ̯ɪrɐ̯ɪʲɐ]	<> বাইরে [bɐ̯ɪre]	কাজ-মাজ [kɐʒ-mɐʒ] বাইরে [bɐ̯ɪre]

Table 4: Examples of dialectal words that are incorrectly transcribed by models. SCB words are underlined.

Dialectal Recall

- Linguists annotate dialectal words in each utterance → ground truth
- Model transcription checked against ground truth:
  Exact match → True Positive (TP)
  Missing or incorrect → False Negative (FN)
- Dialect Recall (per region) = $\frac{\text{TP}}{\text{TP+FN}}$
- Directly measures how well models capture authentic dialect-specific vocabulary

Region	Tugstugi	Tugstugi (Ben-10)	Region	Tugstugi	Tugstugi (Ben-10)
Chittagong	0.033	0.200	Tangail	0.075	0.471
Kishoreganj	0.044	0.156	Habiganj	0.060	0.340
Narsingdi	0.016	0.275	Barishal	0.014	0.400
Narail	0.038	0.325	Sylhet	0.090	0.530
Rangpur	0.033	0.204	Sandwip	0.050	0.342

Acknowledgements

We thank Co-PI Asif Shahriyar Sushmit for his leadership in research design, data curation, and modeling. We acknowledge Brac University, IUT, SUST, Brain Station-23, and APSIS Solutions for their support, and thank Phison, Supermicro, AMD, and the Stanford AI Club for computational resources.