Are ASR foundation models generalized enough to capture features of regional dialects for 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, Bangladesh [2]Brac University, Bangladesh [3]Bengali.AI
tawsiftashwar@iut-dhaka.edu {tanveerazmal, rubayetsabbir, greasat}@gmail.com
md.rezuwan.hassan@g.bracu.ac.bd

Abstract

Conventional research on speech recognition modeling relies on the canonical form for most low-resource languages while automatic speech recognition (ASR) for regional dialects is treated as a fine-tuning task. To investigate the effects of dialectal variations on ASR we develop a 78-hour annotated Bengali Speech-to-Text (STT) corpus named Ben-10. Investigation from linguistic and data-driven perspectives shows that speech foundation models struggle heavily in regional dialect ASR, both in zero-shot and fine-tuned settings. We observe that all deep learning methods struggle to model speech data under dialectal variations but dialect-specific model training alleviates the issue. Our dataset also serves as an out-of-distribution (OOD) resource for ASR modeling under constrained resources in ASR algorithms. The dataset and code developed for this project are publicly available.[1]

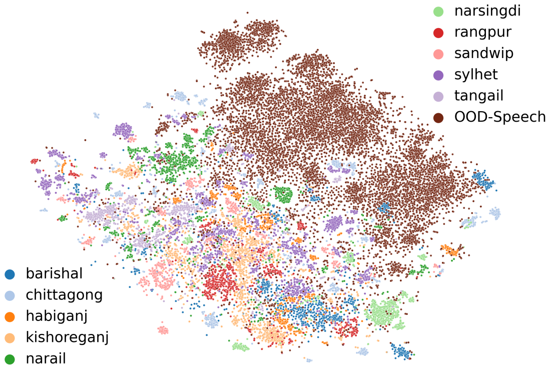


Figure 1: t-SNE plot of GeMAPS features (Eyben et al., 2015) of the standard Bengali speech data (OOD-Speech (Rakib et al., 2023)) and the proposed Ben-10 dataset (10 regions) shows a clear distribution shift.

1 Introduction and Related Works

Computational linguistics has been essential for understanding written and spoken language, aiming to bridge gaps (Wald and Bain, 2008) via natural language interaction in systems. Significant advances in this area encompass automated speech assessment for language learning (Chapelle and Voss, 2016), evaluating language disorders and therapy (Kitzing et al., 2009) and supporting agriculture (Swetha and Srilatha, 2022). Speech-to-Text (STT) is one such technology that is commonly used, yet resources for low-resource languages like Bengali, especially regional variations are lacking. Over the years, a few datasets have been developed for different languages focusing on regional dialects. Researchers in (Shivaprasad and Sadanandam, 2020) created a Telugu dataset for dialect study in speech recognition, (Zhai et al., 2022) focused on the low-resource Yulin dialect in China and (Javed et al., 2024) introduced IndicVoices, a corpus which contains 7k+ hours of everyday dialogues, speeches and readings from 22 Indian languages. Research such as (Hai, 1964) and (Morshed, 1997) focused on modeling standard colloquial Bengali (SCB) phonology, but little has been explored in the diverse regional Bengali dialectal variations in Bangladesh.

Despite advances in the application of deep learning in automatic speech recognition (ASR) (Mehrish et al., 2023) (Kheddar et al., 2024), no prior research has thoroughly investigated their performance in regional Bengali dialects due to the unavailability of diverse datasets. Although existing speech recognition datasets for Bengali (Rakib et al., 2023; Kibria et al., 2022) contain Standard Colloquial Bengali (SCB) and its accents, these lack dialectal variations, which can substantially diverge from

[1] https://github.com/BengaliAI/reg-speech-aacl

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 178–188
December 20-24, 2025 ©2025 Association for Computational Linguistics

SCB. Recent efforts, such as the ChatgaiyyaAlap dataset (Chowdhury et al., 2025), provide resources for converting the Chittagonian dialect to standard Bangla, highlighting the need for dialect-specific data in speech processing. Similarly, BnTTS (Basher et al., 2025) explores few-shot speaker adaptation for Bengali text-to-speech in low-resource settings, underscoring the challenges of adapting models to diverse linguistic contexts. Our research bridges the domain gaps by offering an open-sourced dataset containing spontaneous, diversified regional speech dialects. We demonstrate that available state-of-the-art (SOTA) models struggle to transcribe this dataset in both zero-shot and fine-tuned scenarios. Additionally, we crowdsourced the development of ASR models that outperform existing systems in transcribing speech with regional dialects.

Our contributions presented in this paper are as follows.

1. Introducing the first and largest ASR dataset for 10 regional dialects of Bengali (hence the name **Ben-10**). The dataset contains spontaneous speech data with linguist-validated annotation.

2. Benchmarking of the performance of existing Bengali ASR models, APIs and foundation models on the Ben-10 dataset.

3. Dialect capture rate analysis of the foundational and finetuned models.

2 Regional Dialect Speech Dataset (Ben-10)

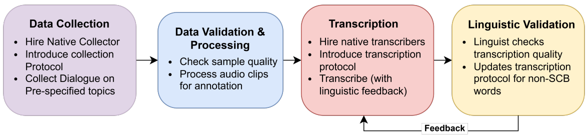

Data Collection
• Hire Native Collector
• Introduce collection Protocol
• Collect Dialogue on Pre-specified topics
Data Validation & Processing
• Check sample quality
• Process audio clips for annotation
Transcription
• Hire native transcribers
• Introduce transcription protocol
• Transcribe (with linguistic feedback)
Linguistic Validation
• Linguist checks transcription quality
• Updates transcription protocol for non-SCB words
Feedback

Figure 2: Ben-10 dataset creation workflow.

We constructed a speech corpus that emphasizes regional dialects of Bengali, drawing from ten distinct regions: Rangpur, Kishoreganj, Narail, Chittagong, Narsingdi, Tangail, Barishal, Habiganj, Sylhet, and Sandwip. Table 1 presents the proportional distribution of data across these regions. To visualize the high-dimensional embeddings in the Geneva feature space, we generated a t-SNE (Van der Maaten and Hinton, 2008) projection into 2D. As shown in Fig. 1, the t-SNE plot reveals a pronounced distribution shift between standard Bengali speech and our Ben-10 dataset, primarily attributable to prosodic variations across dialects and the inclusion of spontaneous speech in the latter. More details on geneva features and comprehensive analysis can be found in the Appendix sections A.1 and A.4. We have illustrated the complete workflow in Fig. 2.

We have collected more than 5 hours of data per region from 394 unique speakers. The responses contain 155 topic stimuli such as family, religion, sports, politics etc. After excluding low-quality samples, recordings were segmented using Silero VAD (Team, 2021) with a maximum segment duration of 30 seconds. Since segmentation boundaries were determined by speech activity rather than fixed intervals, the resulting clips averaged 16.60 seconds in length. The clips were transcribed by 34 trained annotators over 14 months, while linguists validated accuracy. The Ben-10 corpus comprises 16,690 audio clips amounting to 78+ hours of mostly spontaneous speech, 131.38 words per minute and 62,762 unique words. The dataset includes 40.12 hours from males, 32.55 hours from females and 8.35 hours from multiple-gender speakers. The region-wise gender distribution of speakers is shown in Fig. 3.

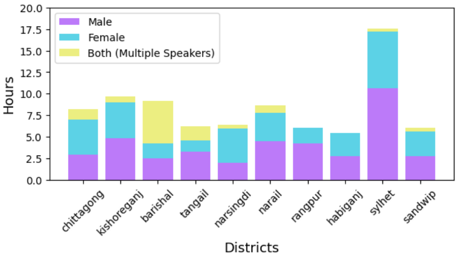


Figure 3: Gender distribution

The dataset was split into train, validation and test sets following an 80:10:10 ratio for each of the dialects. The dialect-wise recording and text length distribution can be found in the Appendix A.2. More details about the dataset can be found in the Appendix A.5

We conducted a NISQA (Mittag et al., 2021) quality evaluation on the speech data. The extracted NISQA features include Mean Opinion Score (MOS), Noisiness, Coloration, Discontinuity and Loudness. From Fig.4 we see that the Discontinuity for Narsingdi is really high, This suggests that speech from this region may have more interruptions, glitches or sudden changes. The audio quality for Narsingdi also suffers from the busier acoustic environment which is indicated by the high Loudness metric. Rangpur also has high Loudness and Discontinuity. More analysis on this can be found in the Appendix A.3.

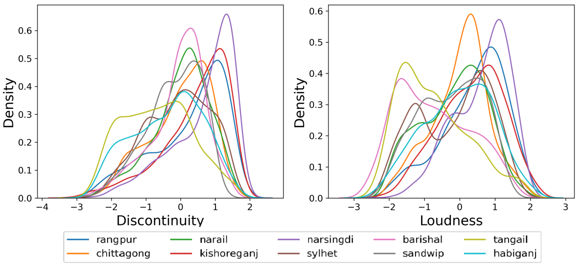


Figure 4: Comparison in the distribution of NISQA features across regions in Ben-10.

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 1: Ben-10 dataset statistics. OOV → words unique to the district that are **O**ut **O**f **V**ocabulary in comparison to SCB (Rakib et al., 2023). M → Male, F → Female, B → Both

3 Dialectic Features of Regional Bengali

Comparing Bengali dialects with SCB reveals notable deviations in all linguistic aspects. Phonetic differences, such as vowel and consonant shifts, morphological variations like unique verb conjugations, syntactic peculiarities including sentence structure and lexical distinctions with region-specific vocabulary highlight the rich tapestry of linguistic diversity. Below, we discuss the intricacies with the International Phonetic Alphabet (IPA) transcript (Fatema et al., 2024).

Phonetic feature analysis of the dialects from the ten regions reveals notable regional variations in vowel and consonant pronunciation. For instance, some unaspirated sounds in Standard Bengali become aspirated in different dialects, especially, the *Sylheti* dialect: unaspirated ক /k/ is pronounced as aspirated 'খ' /k^{h}/ and sometimes as velar fricative খ /x/. Similarly, **Chittagong** dialect is characterized by the change in bilabial plosive sounds প /p/ and ক /k/ into the aspirated ফ /f/ and খ /k^{ɦ}/ sound at the initial syllable position- often at the medial and final position too. One example of this is মাপ/mɐp/ > মাফ/mɐf [English: Measure]. *Sandwip* and *Chittagonian* speakers frequently employ nasalization, particularly in vowel sounds, which contributes to the unique phonetic identity, which is not a characteristic behaviour of the standard colloquial Bengali, e.g. আমার [my] > আঁর [/ɐmɐr/ > /ɐ̃r/]. In *Rangpuri* dialect, it is common to find aspirated sounds /k^{h}/, /t^{h}/, /b^{h}/, /g^{h}/, /d^{h}/, /p^{h}/ being changed into non-aspirated sounds /k/, /d/, /b/, /g/, /p/. *Barishal* speakers exhibit phonetic traits influenced by the eastern Bengali dialect. In *Tangail* dialect, the intonation patterns tend to rise at the end of declarative sentences, giving an inquisitive tone, which contrasts with the typically falling intonation in Standard Bengali. Distinctive prosodic patterns in the *Rangpur* dialect include the frequent use of pitch accents, where certain syllables are emphasized with a higher pitch, affecting the overall rhythm of the speech. **Morphological** feature analysis indicates distinctive word formation and inflectional patterns across the dialects. For example, we find some distinctions in singular and plural inflections in the *Narsingdi* dialect. For example, singular inflection 'টা' /tɐ/ and plural inflection 'গুলা' /gulɐ/ in SCB becomes 'ডা' and 'ডি' /dɪ/, respectively. *Chittagong* dialect speakers exhibit unique verb conjugations and pronouns (e.g. 'আমি করি' /ɐmɪ/ /korɾɪ/ [i do] > 'আঁই গরি' /ɐ̃ɪ/ /goɾɪ/). Different dialects also exhibit high gemination characteristics adding to the lexical richness. For example, in *Barishal* dialect, we have seen the word 'খাইত্তারে' [can eat] /k^{h}ɐɪt̪t̪ɐɾe/ analogous to the SCB verb form 'খেতে পারে' /k^{h}et̪e/ /pɐɾe/. **Syntactic** feature analysis shows deviations from SCB where a sentence follows the subject-object-verb sequence. In the *Chittagong* dialect, the negation word often comes before the verb, which is different from SCB (e.g. 'আমি যাই না (negation)' [I don't go] /ɐmɪ/ /ɟɐi̯/ /nɐ/ > 'আই ন (negation) যাই' /ɐɪ/ /nɔ/ /ɟɐi̯/). Some representative samples of diversity is shown in Table 2 where we can see the high linguistic variations compared to SCB.

4 Benchmarking

We rigorously evaluated transcription capability using pretrained and finetuned foundational ASR models (Table 3). The evaluation was based on word error rate (WER) and character error rate (CER), reported as decimal fractions. In this paper, we refer to large-scale pretrained ASR systems such as Google ASR, Whisper-large and Hishab Conformer as foundational models, since they

Region	Regional Sentence	Standard Bengali	English Translation
Rangpur	মোবাইলোত একনা রেকোড করমো। [mobɐ͜ilot̪ ɜknɐ ɾekod kɔɾmo]	মোবাইলে এখন রেকর্ড করবো। [mobɐ͜ile ɜk^{h}on ɾekɔɾd korbo]	I will record on mobile now.
Kishoreganj	এনো কি লিকছে দ্যাহো। [eno kɪ lɪkc^{h}e d̪ɜho]	এখানে কি লিখেছে দেখো। [ek^{h}ɐne kɪ lɪk^{h}ec^{h}e d̪ɜk^{h}o]	See what is written here.
Narail	তালি তো বাইড়ে গেছে। [t̪ɐlɪ t̪o bɐ͜iɽe gec^{h}e]	তাহলে তো বেড়ে গেছে। [t̪ɐhole t̪o beɽe gec^{h}e]	Then it has increased.
Chittagong	কিসু অইতো ন। [kɪc^{h}u ɔ͡it̪o nɔ]	কিছু হবে না। [kɪc^{h}u hɔbe nɐ]	Nothing will happen.
Sandwip	হেতে আইতো ন [het̪e ɐ͡it̪o nɔ]	সে আসবে না [ʃe ɐʃbe nɐ]	He won’t come.
Sylhet	আমরার লাগিও দোয়া খরিও। [ɐmrɐɾ lɐgi͡o̯ d̪oɐ k^{h}oɾi͡o̯]	আমার জন্যও দোয়া কোরো। [ɐmɐɾ ɟonno͡o̯ d̪oɐ koro]	Pray for me too.
Habiganj	ইস্কুলের কিতা খইতাম? [ɪskuleɾ kɪt̪ɐ k^{h}o͜it̪ɐm?]	স্কুলে কী বলব? [skule kɪ bolbo?]	What to say at school?
Narsingdi	উল্টা-ফাল্টা কতা ক। [ultɐ-p^{h}ɐltɐ kɔt̪ɐ kɔ]	উলটা-পাল্টা কথা বলে। [ultɐ-paltɐ kɔt̪^{h}ɐ bole]	Talk back and forth.
Tangail	এই ছবিডা দ্যাহা অয় নাই। [e͜i c^{h}obɪdɐ d̪ɜha ɔ^{j} na͜i]	এই ছবিটা দেখা হয়নি। [e͜i c^{h}obɪtɐ d̪ɜk^{h}ɐ hɔ^{j}nɪ]	This image has not been viewed.
Barishal	এহন বাড়ি গ্যালে রাগ অয়। [ehon bɐɽɪ gɜle rɐg ɔ^{j}]	এখন বাড়ি গেলে রাগ হয়। [ɜk^{h}on bɐɽɪ gele rɐg hɔ^{j}]	Now I get angry when I go home.

Table 2: Regional dialectal variations in Bengali

are trained on vast amounts of general-purpose multilingual data prior to any task-specific fine-tuning. We used Google’s speech-to-text cloud service API which uses Conformer-based models (Gulati et al., 2020). We investigated the dialect-wise performance of the Wav2vec2 (Baevski et al., 2020) (fine-tuned on SCB (Rakib et al., 2023) and fine-tuned Ben-10). We also evaluated the Hishab Conformer (Nandi et al., 2023) model which was trained on 20k hours of pseudolabelled Bengali speech data. We further benchmarked Whisper-large-v3 (Radford et al., 2023) alongside two models that won previous ASR competitions on SCB (Mashiat et al., 2022) (Addison Howard, 2023): finetuned Wav2Vec2 and Tugstugi (fine-tuned Whisper (Radford et al., 2022)). Foundational models (Google ASR, Hishab, Whisper) performed poorly across all regions. While Wav2Vec2’s performance on SCB is known to be limited (Rakib et al., 2023), fine-tuning did not substantially mitigate the issue. Whisper-based models, particularly Tugstugi, achieved relatively better accuracy but still exhibited high WER and CER. For context, with sufficient training data, these models can achieve WERs below 0.30 in various English benchmarks (Radford et al., 2022).

Several dataset-driven factors appear to underlie these outcomes. Wav2Vec2 struggles particularly in Sandwip and Tangail, as many of these dialectal words do not exist in Standard Bengali (about 62% for Sandwip and 43% for Tangail) and phonetic or prosodic deviations (e.g., Sandwip’s consonant shifts and Tangail’s rising intonation) cause frequent substitutions and deletions. Tugstugi, built on Whisper, benefits from exposure to hundreds of thousands of hours of multilingual data. These diverse pretraining priors appear to enable better adaptation to dialectal variations compared to Wav2Vec2, which remains more aligned with SCB. Moreover, WER reduction varies by dialect due to differences in data volume and linguistic proximity to SCB. Larger splits, such as Sylhet, facilitate better learning despite high divergence, while smaller splits like Barishal show limited gains. Dialects closer to SCB (e.g., Narail, Habiganj) improve more readily than highly distinct ones (e.g., Chittagong, Sandwip). These hypotheses are informed by dataset statistics and linguistic traits; a full causal analysis is planned for future work. More analysis comparing dialect-wise error rates with NISQA metrics can be found in Appendix A.6.

Models	Chittagong		Kishoreganj		Narsingdi		Narail	
	WER	CER	WER	CER	WER	CER	WER	CER
Whisper Large V3	1.09	0.70	1.10	0.81	1.17	0.93	1.48	1.20
Google ASR	1.02	0.98	1.06	1.12	0.87	0.74	1.02	0.94
Hishab Conformer	0.95	0.68	1.02	0.68	0.84	0.56	0.82	0.52
Wav2Vec2 (SCB)	0.97	0.71	0.95	0.79	0.89	0.61	0.87	0.61
Tugstugi	0.93	0.62	0.92	0.72	0.79	0.48	0.76	0.46
Wav2Vec2 (Ben-10)	0.93	0.59	0.92	0.69	0.82	0.54	0.78	0.46
Tugstugi (Ben-10)	**0.85**	**0.45**	**0.86**	**0.52**	**0.64**	**0.29**	**0.62**	**0.29**
Models	Rangpur		Tangail		Habiganj		Barishal	
	WER	CER	WER	CER	WER	CER	WER	CER
Whisper Large V3	1.22	0.99	1.04	0.68	0.91	0.49	1.03	0.72
Google ASR	0.84	0.92	1.25	1.37	1.10	1.01	1.23	1.50
Hishab Conformer	0.88	0.56	0.56	0.29	0.87	0.52	0.87	0.62
Wav2Vec2 (SCB)	0.92	0.71	0.59	0.26	0.88	0.55	0.91	0.67
Tugstugi	0.83	0.56	0.49	0.21	0.79	0.43	0.81	0.54
Wav2Vec2 (Ben-10)	0.81	0.53	0.60	0.30	0.84	0.49	0.84	0.58
Tugstugi (Ben-10)	**0.78**	**0.41**	**0.36**	**0.12**	**0.65**	**0.29**	**0.67**	**0.34**
Models	Sylhet		Sandwip		Average			
	WER	CER	WER	CER	WER	CER		
Whisper Large V3	1.10	0.86	1.14	0.78	1.13	0.81		
Google ASR	0.91	0.78	1.00	0.94	1.03	1.03		
Hishab Conformer	0.90	0.52	0.97	0.67	0.87	0.56		
Wav2Vec2 (SCB)	0.92	0.61	1.00	1.00	0.89	0.62		
Tugstugi	0.87	0.49	0.92	0.66	0.81	0.52		
Wav2Vec2 (Ben-10)	0.85	0.53	1.05	1.01	0.83	0.53		
Tugstugi (Ben-10)	**0.75**	**0.43**	**0.82**	**0.48**	**0.70**	**0.36**		

Table 3: Benchmark results of pretrained and finetuned foundational models. Finetuned versions indicate their respective training datasets in parentheses. For each region, the best WER and CER are in bold.

5 Dialect Transcription Analysis

In this section, we present a method to determine the average rate of successful dialectal word transcription per region. This requires building a vocabulary of region-specific dialectal words, annotating the words in each transcription using this vocabulary, and comparing ASR predictions from multiple models against these annotations to assess whether dialectal variations are accurately transcribed. However, this process demands extensive manual effort, which is beyond the scope of this paper. Therefore, we conduct the experiment on a smaller scale.

Sample transcriptions of 5–10 sentences per region, each containing approximately 50 dialectal words, were extracted by the linguist. Then specific dialectal words were annotated from the transcriptions for evaluating the performance of the models. Then we manually compared the ground truth dialectal words with the transcriptions of the models. If the dialectal word is perfectly transcribed, it is considered a true positive (TP), or else a false positive (FP). For each sentence, we compute the dialect recall rate (i.e., TP/(TP+FP)) for each model (Table 4). Since manual annotation is time-consuming, we only constrain the experiment to the best performing foundational models: Tugstugi and Tugstugi (Ben-10).

The analysis shows that the tugstugi-base model which was not fine-tuned in the regional dialects performed worst across all the regions. The best performing regions were Sylhet and Tangail and the worst were Chittagong and Kishoreganj.

In Table 5, we present some example dialects that are prevalent in their corresponding districts but they are inaccurately transcribed by the models. The foundational model (Tugstugi) often does not predict any text (<> for null prediction) for the dialect word due to the lack of dialect vocabulary in its training corpus. Additionally, foundational models have a strong tendency to default to standard Bengali (Table 4 underlined words). This is expected as the Tugstugi model which is only trained on SCB has no vocabulary for dialectal words and the inferences get aligned with the closest phonetically similar standard Bengali word. The regionally fine-tuned version of the model also does not perform well, but the attempt to map phoneme to grapheme is visible, although the formed words do not match the dialectal annotation.

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

Table 4: Variation of dialect recall observed for different models across different regions.

6 Conclusion

Our study delves into Bengali speech intricacies, creating a 78-hour regional corpus, the sole open ASR dataset for these variations. We find significant morphological and syntactic deviations of different dialects of the SCB, which have not been extensively documented before. This work highlights the challenges SOTA models face in transcribing regional speech, particularly in zero-shot and fine-tuned settings due to limited dialect coverage in the training set. This dataset also provides the first ever regional Bengali text corpus which can be further analyzed to build region centric vocabulary which is crucial in building efficient tokenizers (an integral part of speech encoders).

Region	Ground Truth	Tugstugi	Tugstugi (Ben-10)
Rangpur	ধইরছং [d̪oɪ̯rc^{h}oŋ]	দুঃসা [d̪uʃʃɐ]	দিতো [d̪ɪt̪o]
	গেসনু [gesnu]	<>	গেছেনু [gec^{h}enu]
Kishoreganj	কইতাম [koɪ̯t̪ɐm]	<>	করোনি [kɔronɪ]
	দুইন্নার [d̪uɪ̯nnɐr]	দুনিয়ায় [d̪unɪɐ^{j}]	দুইন্যের [d̪uɪ̯nner]
Narail	মাইসোতো [mɐɪ̯ʃot̪o]	মায়ের সেতু [mɐer ʃet̪u]	মাইনষে ও তো [mɐɪ̯nʃe o t̪o]
	ম্যালাডিক [mɛlɐdɪk]	মেলাডি [mɛlɐdɪ]	মেলাডি [mɛlɐdɪ]
Chittagong	বেককুনে [bekkune]	এখনো [ɛk^{h}ono]	একখোনো [ɛkk^{h}ono]
	হারাপ [hɐrɐp]	<>	খালা [k^{h}ɐlɐ]
Sandwip	লাড়িয়ালান [lɐɽɪɐlɐn]	রিলান্দার [rɪlɐnd̪ɐr]	হেলার [helɐr]
	এগগেরে [eggere]	<>	এগেরে [egere]
Sylhet	ছোখো [c^{h}ok^{h}o]	সৌকে [ʃou̯ke]	ছো ওকে [c^{h} oke]
	তাইনের [t̪ɐɪner]	তাই [t̪ɐɪ̯]	তাই [t̪ɐɪ̯]
Habiganj	আছিলায় [ac^{h}ila^{j}]	চাইছিলা [cɐɪ̯c^{h}ɪlɐ]	চাইছিলায় [cɐɪ̯c^{h}ɪlɐ^{j}]
	খিছুদিন [k^{h}ɪc^{h}ud̪ɪn]	পৃথিবীর [prɪt̪^{h}ɪbɪr]	ফিতুদিন [fɪt̪udɪn]
Narsingdi	খেইলাই [k^{h}eɪ̯lɐɪ̯]	ফেলে [fele]	ফেলায় [fɛlɐ^{j}]
	ডাহাত [dɐhɐt̪]	ডাকতেছে [dɐkt̪ec^{h}e]	ডাত [dɐt̪]
Tangail	দ্যাহে [d̪ɛhe]	দেহে [dehe]	দেহে [dehe]
	থোয় [t̪^{h}o^{j}]	হয় [ho^{j}]	দেখায় [d̪ɛk^{h}ɐ^{j}]
Barishal	কাইলগো [kɐɪ̯lgo]	<>	কাজ-মাজ [kɐɟ-mɐɟ]
	বাইরাইয়া [bɐɪ̯rɐɪ̯^{j}ɐ]	বাইরে [bɐɪ̯re]	বাইরে [bɐɪ̯re]

Table 5: Examples of dialectal words that are incorrectly transcribed by all the models. If the word is transcribed as a vocabulary of SCB it is underlined.

7 Limitations

The lack of computational resources needed to run these DL models is ubiquitous in the Global South and has been plaguing the development of any low-resource language technology development. We alleviated the issue by crowdsourcing finetuning experiments. Future iterations will aim to expand dialectal representation, address gender bias, enhance sentence diversity and refine linguistic analysis to improve understanding of Bengali's regional variations.

8 Ethical considerations

Authors of this paper acknowledge that this research complies with ACL ethical guidelines. This paper is original and any prior work that was used in this research is properly cited. All the findings presented in the paper are truthful and accurate to the best of our knowledge. Data was collected through informed consents and complies with ethical standards. Payments were duly made to those who collected the data (BDT 300 for every hour of speech data collected) and to those who transcribed them (BDT 2000 for every hour of speech data).

9 Acknowledgements

Asif Shahriyar Sushmit, as a Co-Principal Investigator (Co-PI) of the project, made substantial contributions to both the research and the preparation of this manuscript. In particular, he played a key role in problem formulation, data collection and curation, model development, result evaluation, and overall project coordination.

We gratefully acknowledge Brac University, the Islamic University of Technology, Shahjalal University of Science and Technology, Brain Station-23, and APSIS Solutions Ltd. for their financial and logistical support throughout this research. We also extend our appreciation to Phison, Supermicro, AMD, and the Stanford AI Club for providing the computational resources that enabled this work.

References

Ahmed Imtiaz Humayun Addison Howard. 2023. Bengali.ai speech recognition.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Mohammad Jahid Ibna Basher, Md. Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, S. A. Chowdhury, Firoj Alam, Niloofar Yousefi, and Ozlem Ozmen Garibay. 2025. Bntts: Few-shot speaker adaptation in low-resource setting. In *North American Chapter of the Association for Computational Linguistics*.

Carol A Chapelle and Erik Voss. 2016. 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, 20(2):116–128.

Sinthia Chowdhury, Deawan Rakin Ahamed Remal, Syed Tangim Pasha, Ashraful Islam, and Sheak Rashed Haider Noori. 2025. Chatgaiyyaalap: A dataset for conversion from chittagonian dialect to standard bangla. *Data in Brief*, 59:111413.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and 1 others. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Kanij Fatema, Fazle Dawood Haider, Nirzona Ferdousi Turpa, Tanveer Azmal, Sourav Ahmed, Navid Hasan, Mohammad Akhlaqur Rahman, Biplab Kumar Sarkar, Afrar Jahin, Md Rezuwan Hassan, and 1 others. 2024. Ipa transcription of bengali texts. *arXiv preprint arXiv:2403.20084*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Abdul Hai. 1964. *Dhwonibijnan O Bangla Dhwonitottwo*, 3rd edition. Bornomichil.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K M, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, and 2 others. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 109:102422.

Shafkat Kibria, Ahnaf Mozib Samin, M Humayon Kobir, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2022. Bangladeshi bangla speech corpus for automatic speech recognition research. *Speech Communication*, 136:84–97.

Peter Kitzing, Andreas Maier, and Viveka Lyberg Åhlander. 2009. Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, 34(2):91–96.

Mashiat, Md Zarif Ul Alam, Md. Shahrin Nakkhatra, Mobassir, Najibul Haque Sarker, Ramisa Alam, Sheikh Azizul Hakim, Sushmit, Tahsin, and Zaber Ibn Abdul Hakim. 2022. Dl sprint - buet cse fest 2022. https://kaggle.com/competitions/dlsprint. Kaggle.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869. Version 3.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

Abul Kalam Manzur Morshed. 1997. *Adhunik Bhashatatwa*, 2nd edition. Noya Udyog.

Rabindra Nath Nandi, Mehadi Hasan Menon, Tareq Al Muntasir, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Tariqul Islam, Shammur Absar Chowdhury, and Firoj Alam. 2023. Pseudo-labeling for domain-agnostic bangla automatic speech recognition. *arXiv preprint arXiv:2311.03196*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Fazle Rabbi Rakib, Souhardya Saha Dip, Samiul Alam, Nazia Tasnim, Md. Istiak Hossain Shihab, Md. Nazmuddoha Ansary, Syed Mobassir Hossen, Marsia Haque Meghla, Mamunur Mamun, Farig Sadeque, Sayma Sultana Chowdhury, Tahsin Reasat, Asif Sushmit, and Ahmed Imtiaz Humayun. 2023. Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking. *Preprint*, arXiv:2305.09688.

S Shivaprasad and M Sadanandam. 2020. Identification of regional dialects of telugu language using text independent speech processing models. *International Journal of Speech Technology*, 23:251–258.

Polisetty Swetha and Jenega Srilatha. 2022. Applications of speech recognition in the agriculture sector: A review. *ECS Transactions*, 107(1):19377.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6:435–447.

MengEn Zhai, LiHong Dong, Yi Qin, and FeiFan Yu. 2022. The research of chain model based on cnn-tdnnf in yulin dialect speech recognition. In *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, pages 883–888.

A Appendix

A.1 Geneva Features

Definitions for the Geneva features mentioned in the paper are provided below:

- **F3frequency_sma3nz_amean:** Average frequency of the third formant
- **F3amplitudeLogRelF0_sma3nz_stddev-Norm:** Normalized standard deviation of the logarithmic difference between the amplitude of the third formant and the fundamental frequency
- **alphaRatioV_sma3nz_stddev-Norm:** Normalized standard deviation of alpha ratio in voiced segments
- **hammarbergIndexV_sma3nz_stddev-Norm:** Normalized standard deviation of Hammarberg index in voiced segments
- **slopeV0-500_sma3nz_stddev-Norm:** Normalized standard deviation of slope in the frequency range 0-500Hz for voiced segments
- **F0semitoneFrom27.5Hz_sma3nz_amean:** Average fundamental frequency in a logarithmic scale - a frequency scale with semitones starting at 27.5Hz (semitone 0)
- **F0semitoneFrom27.5Hz_sma3nz_stddev-Norm:** Normalized standard deviation of fundamental frequency in semitones starting at 27.5Hz
- **F0semitoneFrom27.5Hz_sma3nz_percentile20.0:** 20th percentile of fundamental frequency in semitones starting at 27.5Hz
- **F0semitoneFrom27.5Hz_sma3nz_percentile50.0:** Median (50th percentile) of fundamental frequency in semitones starting at 27.5Hz
- **loudness_sma3_pctlrange0-2:** Range of loudness within the 0-2 percentile
- **mfcc1_sma3_amean:** Average of the first Mel-Frequency-Cepstral-Coefficients
- **alphaRatioV_sma3nz_amean:** Average alpha ratio (ratio of the summed energy from 50–1000 Hz and 1–5 kHz) in voiced segments
- **hammarbergIndexV_sma3nz_amean:** Average Hammarberg index (ratio of the strongest energy peak in the 0–2 kHz region to the 2–5 kHz region) in voice.

A.2 Dialect-wise recording and text length distribution

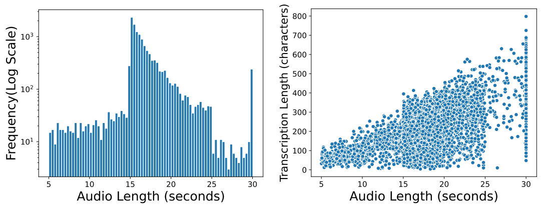


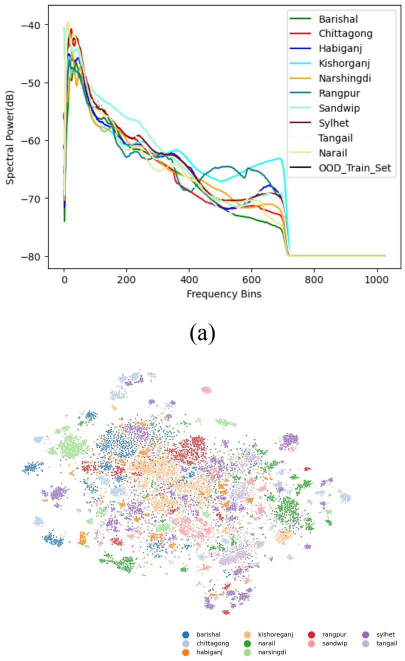
Figure 5: (Left) Audio length distribution and (Right) Scatter plot showing the relationship between Audio length and transcript length

Fig. 5 shows that most of the recordings are around 15 seconds in length and none are above 30 seconds. Furthermore, from Fig. 5 we observe that the recording length and transcript character count are not strictly correlated. Calculating the long-term spectral average for different dialects, we obtain Fig. 6(a). Here we see dialects diverge both semantically and spectrally. This demonstrates the high distribution shift among dialects. As we ensured the same recording protocol, this distribution shift is suspected to be caused by the prosodic variations present in the different dialects.

A.3 Recording Quality Evaluation using NISQA metrics

We performed a NISQA (Mittag et al., 2021) quality evaluation of the speech data. The features calculated are Mean Opinion Score, Noisiness, Coloration, Discontinuity and Loudness. Definitions of these features can be found in the NISQA paper. The distribution difference of SCB OOD-Speech and Regional Ben10 Datasets are provided in Fig. 9. Noisiness is worse in our dataset when compared to OOD-Speech due to recording in an uncontrolled setting with the presence of environmental noise. Our corpus fares well in Discontinuity and is similar in Coloration compared to OOD-Speech. Since the audio is recorded with a phone mic and it is often not possible to hold the phone at the speakers mouth Loudness is worse.There is no visible difference in audio processing which results in a similar Coloration.

MOS (Mean Opinion Score): MOS is a subjective speech quality evaluation metric based on the



(a)

(b)

Figure 6: (Left) Long-term Spectral Average of Recordings from different dialects (Right) t-SNE plot of Geneva features of 10 dialects in Ben10 dataset

opinions of a diverse group of listeners, averaging their ratings on a scale from 1 to 5 (poor to excellent). It incorporates aspects such as clarity, distortion, and intelligibility and is influenced by the crowd's diversity and listening conditions.

Noisiness: Noisiness is a metric to measure the amount of unwanted background noise in an audio signal. Excessive noise can overwhelm the original speech, making it harder to listen to and understand. It is often used to calculate the signal-to-noise ratio and how different types of noise affect the overall audio quality.

Coloration: Coloration measures alterations to the speech signal, often due to audio processing or transmission. This can include changes in the frequency spectrum such as addition, removal, shift, amplification, or attenuation in certain frequency bands resulting in a speech sound that may appear unnatural or distorted.

Discontinuity: Discontinuity evaluates sudden interruptions or abrupt changes in the speech signal such as gaps, glitches, or shifts in audio levels. These can disrupt the smooth flow of speech and decrease its perceived quality.

Loudness: Loudness measures the perceived volume of the speech signal. Proper loudness ensures the speech is audible without being overwhelming or distorted.

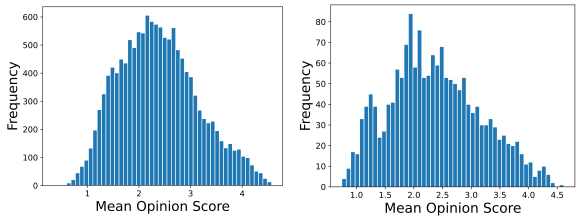


Figure 7: Mean Opinion Score(MOS) Distribution of the Train and Validation fold.

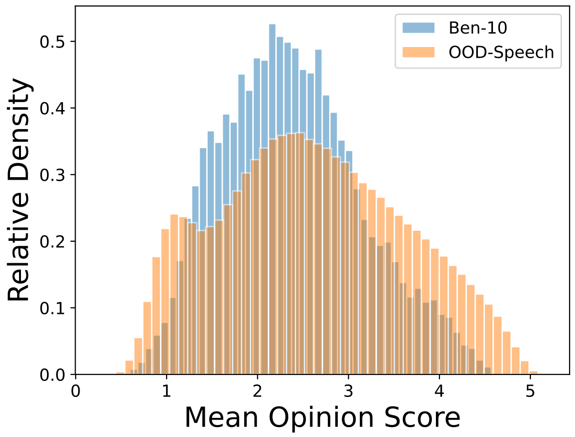


Figure 8: MOS comparsion between ben10 and OOD speech

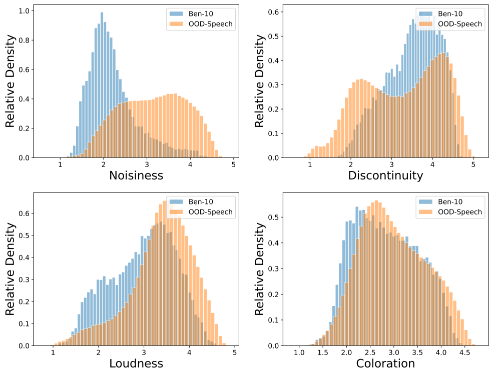


Figure 9: Comparison in distribution of NISQA features in Ben-10 🟦 and OOD-Speech 🟧 Datasets

A.4 Geneva feature Analysis

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (Eyben et al., 2015) provides a compact yet informative representation of speech signals. It is a set of acoustic features used to characterize various aspects of speech signals, such as pitch-related features, energy-related features, timing-related features,

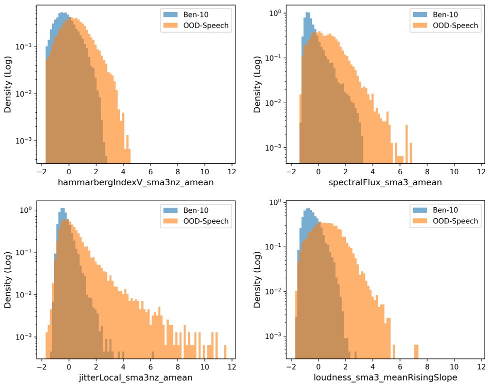


Figure 10: Smile feature histogram comparisons between Ben-10 and OOD-Speech Datasets

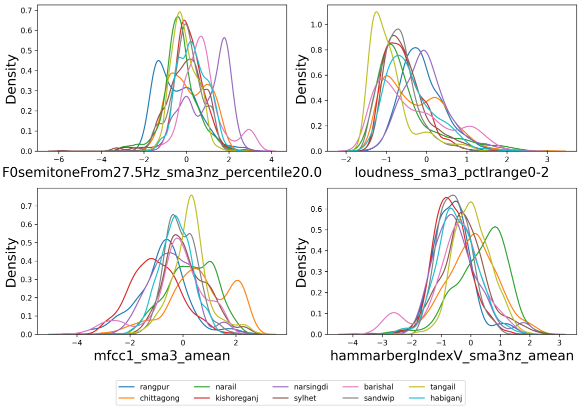


Figure 11: KDE plot of several geneva features showing shifted distributions for different dialects from the Ben-10 Dataset.

spectral features, and voice-quality related features. **slopeV0-500_sma3nz_stddevNorm** denotes the Normalized standard deviation of slope in the frequency range 0-500 Hz for voiced segments. **F3amplitudeLogRelF0_sma3nz_stddevNorm** denotes the Normalized standard deviation of the logarithmic difference between the amplitude of the third formant and the fundamental frequency. These parameters are related to the emotional state and voice control. Less spread (OOD-Speech) indicates a steadier or more controlled voice, while greater variability (Ben10) suggests emotional arousal or less control (as often observed in colloquial speech). **F0semitoneFrom27.5Hz_sma3nz_amean** refers to voice pitch distribution. A balanced dataset (male and female) will have a bimodal distribution which can be seen in Ben10 but the bi-modal peaks are not clearly observable in OOD-Speech as it is male-heavy.

A.4.1 Feature Histograms

Fig. 10 shows observable distribution shifts in some individual Geneva features for the two sources. The OOD-Speech dataset showcases broader distributions, whereas the Ben-10 dataset often shows higher density in the lower values. For metrics such as slopeV0-500_sma3nz_stddevNorm, the Ben-10 dataset concentrates on the lower part of the scale signifying low slope deviation from the mean. Additionally, the distribution of F0semitoneFrom27.5Hz_sma3nz_percentile20.0 is right-skewed in both datasets.

A.4.2 KDE Plot

We analyzed KDE plots of various Geneva features across different dialects in Fig. 11. The kernel Density Estimation (KDE) plots compare the distribution of various features across different district samples. Most districts show a central tendency, indicating similar behavior, while Habiganj, Narsingdi and Tangail often deviate from the central concentration. This suggests greater variability and dispersion in their acoustic features.

A.5 Dataset

- Data gathering followed specific protocols to encourage natural speech from the aforementioned regions.

- Conversation topics varied from family, university, hostel, sports, religion, profession, studies, gossip, childhood, news, politics, agriculture, daily life etc. All conversations were informal to ensure the inclusion of authentic regional dialects.

- After data collection, we cleaned the data by excluding samples that did not follow the protocol or had poor acoustic quality.

- The audios were uploaded onto the Labelbox data annotation platform. A total of 34 local transcribers, trained by a linguist were engaged in transcription which took approximately fourteen months. Residents from the specific region where the data was collected were employed as transcribers based on their performance in a linguist-evaluated transcription test involving 100 audio samples from that region. While performing **validation**, an expert linguist evaluated spelling and transcription errors to ensure homogeneity and correctness of the data and also provided feedback to the annotators.

A.6 Result Analysis

Fig. 12 and Fig. 13 show the district-wise correlation (Pearson coefficient) between WER/CER and NISQA metrics for our benchmark Whisper (Medium) model. Higher absolute values indicate stronger linear relationships, with positive coefficients reflecting direct proportionality and negative ones indicating inverse trends.

In Fig. 12, **Narsingdi** exhibits the strongest positive correlation between WER and coloration, discontinuity and loudness, suggesting that perceived audio distortions and loudness significantly drive transcription errors in this dialect.

Conversely, **Barishal** displays strong *negative* correlations, implying that higher perceived quality in these dimensions corresponds to *lower* error, highlighting its relative robustness among tested dialects.

These findings underscore that dialect-specific acoustic characteristics interact differently with perceived quality dimensions, and NISQA metrics can serve as useful proxies for predicting ASR reliability in low-resource, dialectally diverse settings like Bengali.

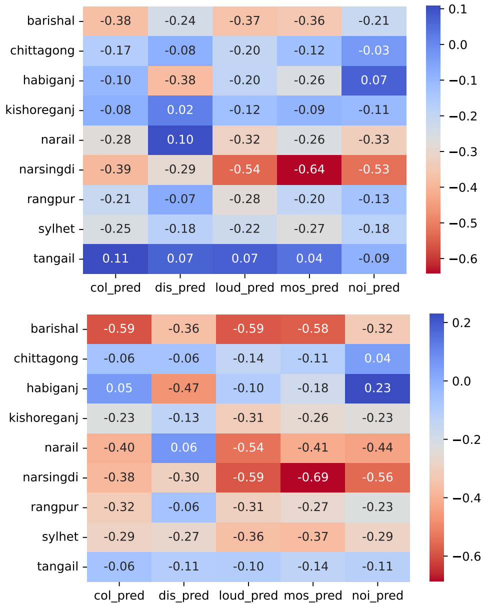


Figure 12: District-wise correlation (Pearson Coefficient) of our benchmark Whisper (Medium) model and NISQA metrics {WER(top) and CER(bottom)}

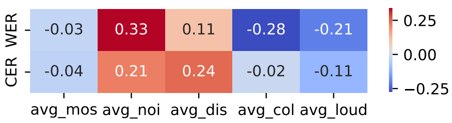


Figure 13: Correlation (Pearson Coefficient) of our benchmark model performance and NISQA metrics.

B Competition reports

We massively crowd-sourced fine-tuning experiments through a public **Bengali Regional Speech Recognition** research competition on Kaggle, which lasted 20 days. The competition resulted in 399 submissions from 83 teams consisting of 247 competitors. More details about the competition can be found on the competition page. The top 11 teams were selected based on the WER score in the publicly available validation set. The public leaderboard (LB) was created using 51% of the validation data while there was another private LB with the rest of the data which was only accessible by competition hosts. Next, these models were evaluated with the hidden test set. The final standing was decided by aggregating the WER scores (validation and test), report and presentation scores. The competitors documented their approach and wrote reports on their pipeline. The pdfs can be found in the following link: Competition Reports