Transcribing Regional Bangladeshi Dialects: A Dual-Stage Sequential Fine-Tuning Approach

Md Nasiat Hasan Fahim, **Miftahul Alam Adib**, **Arif Hussain**
Shahjalal University of Science and Technology
nhfahim18@gmail.com, miftahuladib04@gmail.com, arif1147@gmail.com

Abstract

This paper presents a robust Automatic Speech Recognition (ASR) framework tailored for the linguistically diverse landscape of Bangladesh, developed for the Shobdotori ASR challenge. The primary objective is to accurately transcribe regional dialectal speech from 20 distinct districts—including Chittagong, Sylhet, and Barishal—into Standard Formal Bangla. We propose a transfer learning approach leveraging the OpenAI Whisper (Medium) architecture, optimized via a novel Dual-Stage Sequential Fine-Tuning strategy. To mitigate the scarcity of dialectal data, our curriculum augments the primary dataset of 3,350 samples with 6,108 filtered samples from external corpora. Furthermore, our pipeline incorporates a sequence-to-sequence normalization strategy to address significant domain shifts caused by phonetic, lexical, and prosodic variances. Experimental evaluations demonstrate the system's efficacy, achieving a Normalized Levenshtein Similarity (NLS) score of **0.913** on the public leaderboard and **0.881** on the private leaderboard. These results confirm that our end-to-end methodology effectively bridges the gap between dialectal pronunciation and standardized orthography.

1 Introduction

Bengali (Bangla) is ranked as the sixth or seventh most spoken language globally, boasting over 242 million native speakers and a significant diaspora population (World Population Review, 2025). While Standard Colloquial Bangla serves as the official medium for education and administration in Bangladesh—a nation where 98% of the population speaks the language—the linguistic landscape is characterized by profound internal variation. Linguists classify these variations into four major dialect groups: North Bengal, Rajbanshi/Rangpuri, East Bengal, and South Bengal. The scale of these dialects is massive; for instance, Chittagonian is spoken by approximately 13–16 million people, Sylheti by over 11 million, and Rangpuri by 10 million (Dhaka Tribune, 2024). These varieties differ so significantly from the standard form in phonology, vocabulary, and grammar that some linguists argue they constitute distinct languages within the Indo-Aryan family.

Current state-of-the-art ASR systems are predominantly trained on Standard Bangla, causing performance to degrade severely when exposed to these regional varieties. This "accent mismatch" creates a technological divide that excludes millions of native dialect speakers. The challenge is threefold: acoustic variability, morphological variation, and lexical divergence. Recent phonological studies have documented over 13 distinct variation patterns across 20 regions (Rahman et al., 2024). For example, the standard bilabial stop /p/ frequently shifts to the glottal fricative /h/ in Noakhali (e.g., "pani" becoming "hani"), whereas it shifts to the labiodental fricative /f/ or /ph/ in Sylhet and Chittagong (Rahman et al., 2024). Additionally, morphological deviations, such as unique verb conjugations in Chittagonian and gemination in Barisal, introduce complexity for models relying on standard sub-word tokenization.

To address these linguistic disparities, we propose a solution within the framework of the Shobdotori competition. We initialize our architecture using the 1st place solution checkpoint from the Bengali.AI Speech Recognition competition (Howard et al., 2023), effectively leveraging its established Bengali acoustic priors. By employing a Dual-Stage Sequential Fine-Tuning strategy with Low-Rank Adaptation (LoRA) and auxiliary datasets, we further adapt this model to the specific acoustic nuances of regional dialects. Our work contributes a sequential transfer learning framework that effectively adapts a robust pre-trained Bengali ASR checkpoint to diverse acoustic domains, utilizing high-rank LoRA to capture complex dialectal variations and "long-tail" vocabulary.

2 Related Work

Research in Bengali Automatic Speech Recognition (ASR) has progressed significantly over the last decade. Early approaches primarily utilized Hidden Markov Models (HMMs) for isolated word recognition (Sultana et al., 2021). With the advent of deep learning, focus shifted towards Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Nayem et al., 2023). However, the development of ASR for dialects has historically been impeded by data scarcity.

Recent years (2024–2025) have seen a surge in dialect-

focused research. Chakraborty et al. (Chakraborty et al., 2022) developed CNN-based models specifically for Sylheti. Chowdhury et al. (Chowdhury et al., 2024) introduced *ChatgaiyyaAlap*, a parallel corpus for Chittagonian. More recently, Shawon et al. (Shawon et al., 2025) examined computational processing of spontaneous speech across dialects, while the *BanglaTalk* system (Hasan et al., 2025) targeted real-time dialectal assistance.

The field is rapidly moving towards large-scale benchmarks. The release of datasets like *Vashantor* (Anonymous, 2025d), which benchmarks dialect translation accuracy, and *ONUBAD* (Anonymous, 2025b), a comprehensive dialect translation dataset, highlights the growing interest in this domain. Furthermore, comparative studies on bridging Standard Bangla with regional variants have begun utilizing Transformer models to quantify linguistic distance (Anonymous, 2025a,c). Our work distinguishes itself by targeting the broad spectrum of 20 distinct regional dialects under the Shobdotori challenge, leveraging the large-scale Whisper architecture to perform end-to-end sequence normalization.

3 Methodology

Our approach combines three key components: fine-tuning pre-trained Whisper models on dialectal data, implementing robust data preprocessing strategies, and applying post-processing for transcription refinement. The complete pipeline is visualized in Figure 1.

3.1 Model Architecture

We employed the **Whisper Medium** architecture (769M parameters), a Transformer-based encoder-decoder model trained on 680,000 hours of weakly supervised speech data. The pre-training corpus is highly diverse, consisting of 65% English audio, 18% non-English audio with English transcripts, and 17% non-English audio with native transcripts covering 98 languages. This massive multilingual pre-training provides the model with robust generalization capabilities for diverse acoustic environments.

3.2 Initialization Strategy

Rather than starting from the generic multilingual weights, we initialized our model using the checkpoint from the 1st Place Solution of the Bengali.AI Speech Recognition competition (Howard et al., 2023), which is open-sourced under the CC0: Public Domain license on Kaggle. This initialization provides a superior starting point for Bengali ASR due to its rigorous domain-specific training. The checkpoint was trained on high-performance infrastructure consisting of 8× 48GB RTX A6000 GPUs using the Hugging Face Trainer (batch size 8, learning rate 1e-5) for 50,000 steps. It utilized a comprehensive aggregation of diverse datasets, including OpenSLR (37 & 53), MadASR, Shrutilipi, Macro, Kathbath, and pseudo-labeled YouTube videos. Furthermore, the model was exposed to aggressive augmentation strategies such as spectrogram dithering, time/frequency masking, Libsonic-based speed/pitch perturbation, and resampling (16kHz→8kHz→16kHz). Finally, a custom Bengali-specific tokenizer with a 12k vocabulary was trained to replace the original tokenizer, significantly enhancing inference speed and enabling larger beam widths.

We implemented Low-Rank Adaptation (LoRA) on top of this checkpoint with a high-rank configuration (Rank 1024, Alpha 64, Dropout 0.1) targeting the `q_proj` and `v_proj` modules.

3.3 Data Collection and Preprocessing

To ensure the robustness of our ASR models, we implemented a standardized data processing pipeline. Our training corpus aggregates data from three distinct sources, totaling approximately 9,458 audio samples. The comparative statistics regarding sample count and lexical diversity across these datasets are detailed in Figure 2.

The primary dataset consists of 3,800 audio recordings (3,350 train, 450 test) spanning 20 distinct regional dialects such as Chittagong, Barisal, Sylhet, and Mymensingh. The dataset exhibits significant class imbalance, with regional sample sizes ranging from 21 to 401 files, as shown in Figure 3. This is augmented by Auxiliary Dataset A (DL Sprint) and Auxiliary Dataset B (AI Speech).

Given the crowdsourced nature of the auxiliary datasets, we applied rigorous quality control. We analyzed the temporal characteristics of the audio to determine optimal filtering thresholds (see Figure 4). For DL Sprint, we filtered for sentence lengths between 4 and 11 words. For the AI Speech subset, we retained only concise phrases between 4 and 5 words. We adopted a "Mixed Training, Separate Validation" strategy where training portions were concatenated, but validation sets (Val-Main and Val-Diff) were kept separate.

3.4 Fine-tuning Strategy

Our training pipeline prevents "catastrophic forgetting" via two phases: Phase 1 (Base Adaptation) on Main + DL Sprint, and Phase 2 (Targeted Refinement) on Main + AI Speech. We implemented a Dual-Stream Evaluation mechanism using a unified composite score:

$$S_{final} = (\lambda_{main} \times WER_{main}) + (\lambda_{diff} \times WER_{diff}) \quad (1)$$

By increasing the weight of the Main dataset to 0.95 in Phase 2, we forced the model to prioritize features learned in Phase 1.

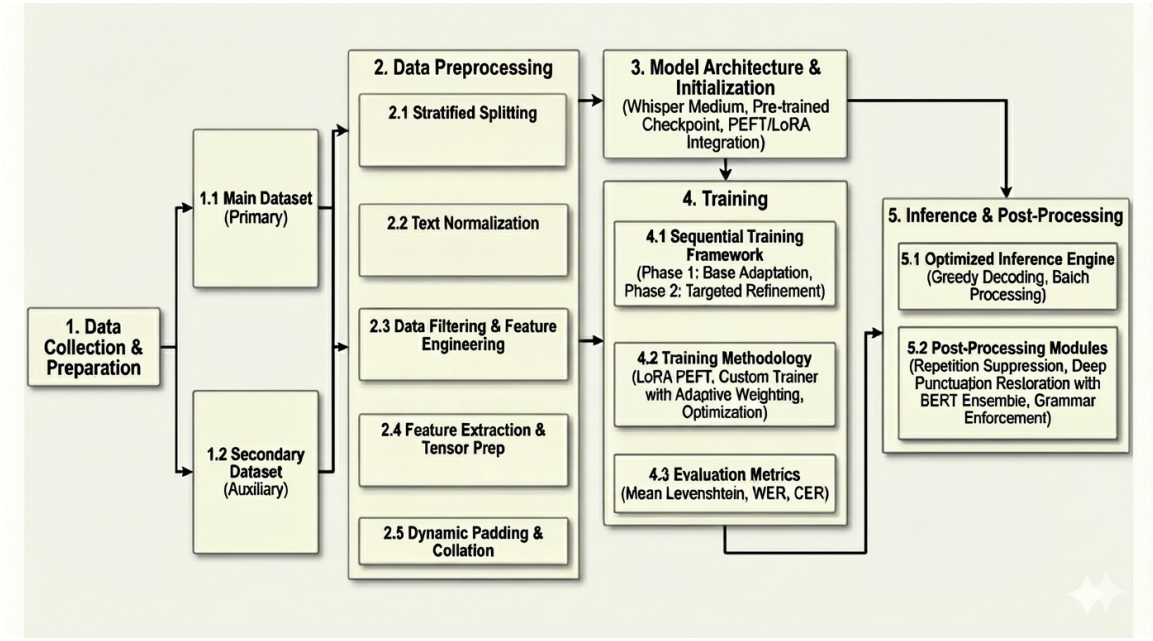

Figure 1: **Proposed Dual-Stage ASR Framework.** The pipeline illustrates the sequential transfer learning from Phase 1 (Base Adaptation) to Phase 2 (Domain Refinement), followed by the BERT-based post-processing module. The "Data Preprocessing" block feeds into a LoRA-integrated Whisper architecture, utilizing adaptive weighting to ensure stability while learning new dialectal features.

Table 1: Comparative Training Configurations

Config	Phase 1	Phase 2	Rationale
Initialization	whisper-med	Ph1 Best	Transfer learning
Weights	0.89 / 0.11	0.95 / 0.05	Prevent forgetting
Epochs	10	8	Prevent overfitting
Warmup	100	0	Stable start
LR	1×10^{-4}	1×10^{-4}	Consistent

3.5 Inference Pipeline

We implemented an optimized inference engine using greedy decoding (`num_beams=1`) and batch processing. Following generation, we applied a specialized post-processing module. This involves repetition suppression by truncating word sequences that repeat more than 8 times, and punctuation restoration using an ensemble of four **MuRIL-base-cased** models—utilizing 6, 8, 11, and 12 layers respectively—trained to predict periods, commas, and question marks. To correct the under-prediction of sentence terminators, we applied a class weight vector favoring the Dari (।).

4 Results and Analysis

4.1 Experimental Setup

All experiments were conducted in a standardized high-performance computing environment using a dual-GPU setup (NVIDIA Tesla T4 x2). Given the VRAM constraints (approx. 15GB per GPU), we employed Mixed Precision (FP16) training and Gradient Accumulation. We architected the training pipeline as a Two-Stage Curriculum Learning process. Experiment A (Base Adaptation) utilizes the merged Main and DL Sprint dataset, while Experiment B (Domain Refinement) applies a stricter regularization strategy using the Main and AI Speech dataset. We optimized Cross-Entropy Loss and monitored performance using CER as a training proxy and WER to ensure semantic integrity.

Table 2: Hardware and Runtime Specifications

Component	Specification
GPU Model	NVIDIA Tesla T4 (x2)
Total VRAM	15,360 MiB per GPU
CUDA Version	12.6
Framework	PyTorch (Hugging Face)

4.2 Evaluation Metric

The competition evaluates submissions using Normalized Levenshtein Similarity (NLS). Calculated as:

$$NLS(P, R) = 1 - \frac{Lev(P, R)}{\max(|P|, |R|)} \tag{2}$$

It rewards phonetically close approximations of complex Bengali conjuncts (Juktakkhor).

4.3 Results

Scaling to Whisper Medium with our pipeline resulted in a significant performance shift. While the Single-Stage

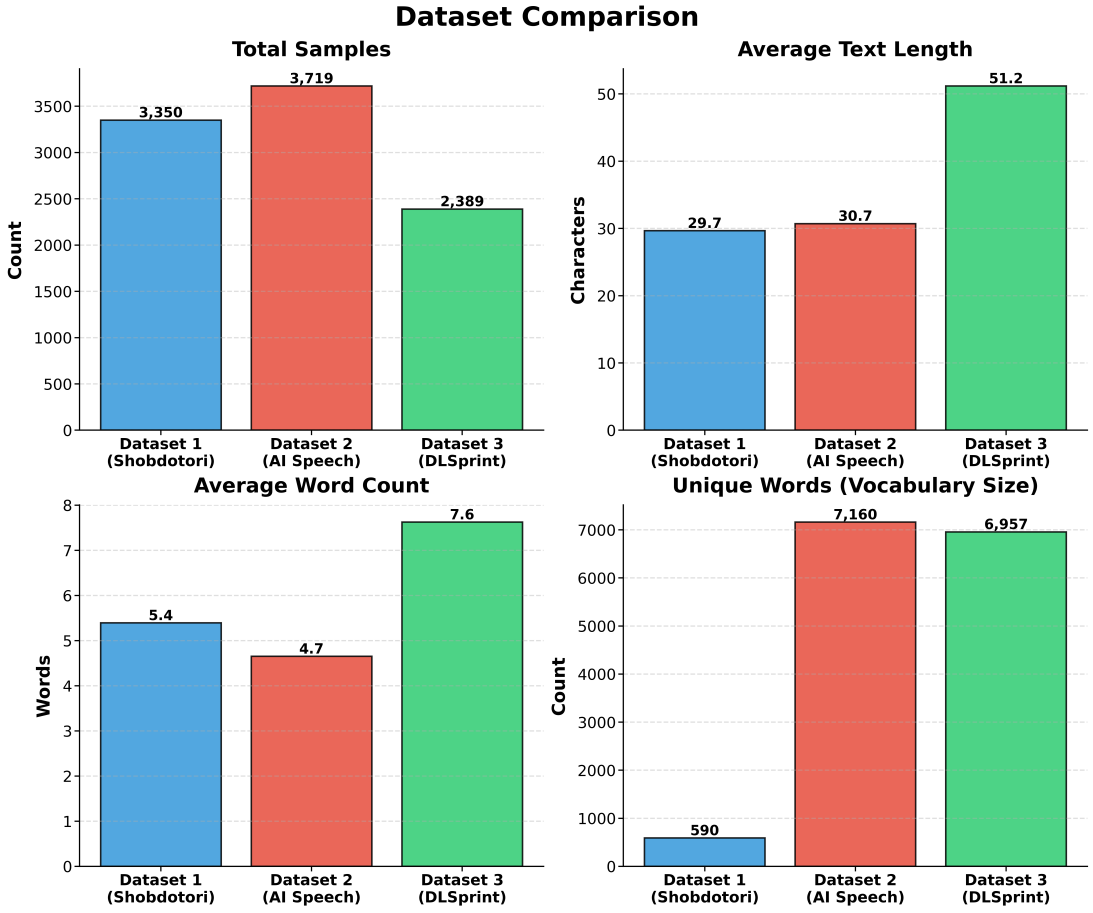

Figure 2: **Dataset Comparison Analysis.** The bar charts illustrate the distribution of Total Samples, Average Text Length, Word Count, and Vocabulary Size (Unique Words) across the three datasets used. While the primary dataset (Shobdotori) provides targeted dialectal data, the auxiliary datasets (AI Speech and DLSprint) significantly contribute to the vocabulary size (approx. 7,000 unique words each) and overall training volume.

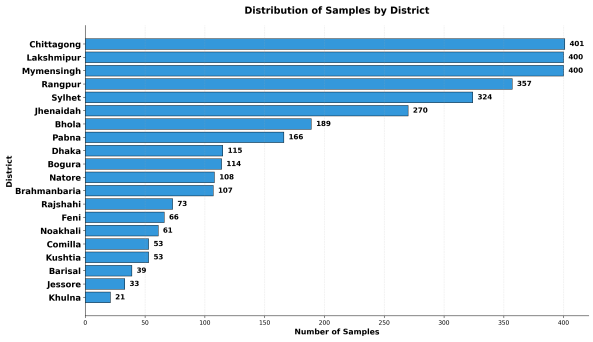

Figure 3: **Geographic Distribution of Dialectal Data.** This visualization highlights the sample density across 20 districts of Bangladesh. A significant class imbalance is observed, with major regions like Chittagong, Lakshmipur, and Mymensingh (approx. 400 samples each) dominating the corpus.

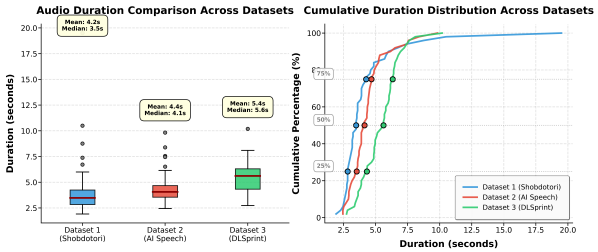

Figure 4: **Audio Duration Analysis.** The box plots (left) indicate that the Mean duration across all datasets hovers between 4.2s and 5.4s. The Cumulative Duration Distribution (right) reveals that over 75% of the Shobdotori and AI Speech datasets consist of clips under 5 seconds.

model achieved a marginally higher Public Leaderboard score, the Proposed Dual-Stage Model achieved a superior Private Leaderboard score of **0.88077**. This confirms that our adaptive weighting strategy successfully improved robustness and out-of-distribution generalization, as seen in Figure 5.

5 Error Analysis

Our approach faced several constraints and challenges. Computationally, we were restricted to a batch size of 4

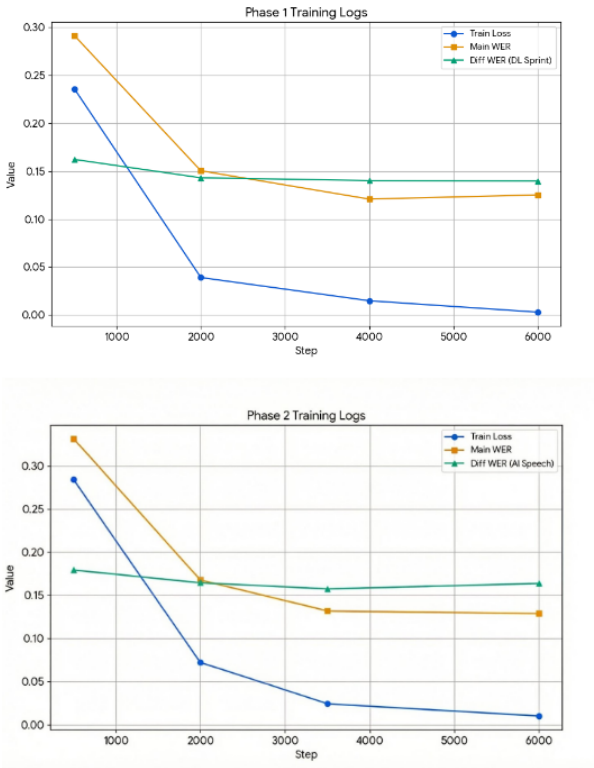

Figure 5: **Training Log Analysis.** The curves demonstrate the effectiveness of Adaptive Weighting across two phases. Phase 1 (top) stabilizes the base model, while Phase 2 (bottom) shows the Main Dataset WER (orange) consistently improving alongside the specific target domain.

Table 3: Comparative Experimental Results

Exp ID	Config	Public NLS	Private NLS
Baseline 1	Static Pad	0.76897	0.71913
Baseline 3	Dynamic Pad	0.83557	0.80765
Interim 2	Main Only	**0.91664**	0.87203
Proposed	**Dual-Stage**	0.91345	**0.88077**

and had to split training into two disjoint notebooks due to 15GB VRAM limits and runtime quotas.

Acoustically, we observed significant variability in phoneme realization which complicated the model's learning process. A primary source of confusion was the acoustic shift where the standard bilabial /p/ (as in *Pani*, meaning water) transforms into the regional fricative /f/ (as in *Fani*). This phonetic deviation often led to misclassification in the early training stages.

Lexically and morphologically, the mapping between Standard and Regional forms proved complex. The model struggled with one-to-many mappings, such as the Standard verb *Jabo* ("I will go") corresponding to multiple regional variants like *Zaiyum* or *Zamu* depending on the specific dialect. Furthermore, semantic ambiguity caused context failures; for instance, the token *Komole* presented a dual challenge: it can function as a proper noun in the locative case ("in Komol") or as a conditional verb form ("if it lessens"). The model occasionally struggled to resolve this context-dependent polysemy without broader semantic cues.

Additionally, the post-processing modules sometimes introduced artifacts, such as premature sentence terminators in complex sentences or incomplete normalization of colloquialisms.

6 Conclusion

In this work, we presented a comprehensive framework for bridging the linguistic divide between Standard Bangla and its diverse regional variations. By leveraging the OpenAI Whisper Medium architecture initialized with a robust, domain-adapted checkpoint, we successfully developed an end-to-end ASR system capable of normalizing speech from 20 distinct Bangladeshi districts. Our approach effectively addresses the "accent mismatch" problem, where traditional models fail due to significant phonetic deviations (such as the /p/ to /h/ or /f/ shifts) and morphological complexities inherent in dialects like Chittagonian and Sylheti.

The core of our success lies in the *Dual-Stage Sequential Fine-Tuning* strategy coupled with *Adaptive Weighting*. This curriculum learning approach allowed the model to first stabilize on a broad mixture of dialectal data before refining its sensitivity to specific domain shifts in the second phase. Crucially, the adaptive weighting mechanism prevented catastrophic forgetting—a common pitfall in transfer learning—by enforcing a strong regularization penalty based on the primary dataset's performance. Furthermore, the application of High-Rank Low-Rank Adaptation (LoRA) provided the necessary plasticity to capture "long-tail" dialectal vocabulary without the prohibitive cost of full-parameter fine-tuning. The resulting NLS score of **0.88077** on the private leaderboard attests to the system's ability to generalize to unseen speakers and regional variances.

Looking forward, we identify two critical avenues for advancement. First, while our acoustic model performs implicit normalization, integrating a dedicated sequence-to-sequence Large Language Model (such as BanglaT5) in the post-processing stage could explicitly handle complex grammatical restructuring, shifting the burden of "translation" from the acoustic encoder to a semantic decoder. Second, to address the severe class imbalance observed in districts like Narail and Meherpur, we plan to employ synthetic data generation via Text-to-Speech (TTS) and voice conversion technologies, creating a more phonetically balanced training corpus for future iterations.

Acknowledgments

We thank the AI-FICATION organizing committee and the Department of Electronics & Telecommunication

Engineering, CUET, for organizing this competition.

References

Anonymous, "Bridging Dialects: Translating Standard Bangla to Regional Variants," *arXiv preprint arXiv:2501.05749*, 2025.

Anonymous, "ONUBAD: A comprehensive dataset for Bangla dialect translation," *ScienceDirect Data Article*, 2025.

Anonymous, "Quantifying Linguistic Variation in Bangla through Dialect-to-Dialect Translation," *IEEE Xplore*, 2025.

Anonymous, "Vashantor: A Large-Scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects," *SSRN*, 2025.

G. Chakraborty et al., "Soft-computation based speech recognition system for Sylheti language," *Int. J. Speech Technol.*, vol. 25, 2022.

S. Chowdhury et al., "ChatgaiyyaAlap: A dataset for conversion from Chittagonian dialect to standard Bangla," *Data in Brief*, 2024.

Dhaka Tribune, "Chittagonian, Sylheti ranked among 100 most spoken languages," 2024.

M. Hasan et al., "BanglaTalk: Towards real-time speech assistance for Bengali regional dialects," *arXiv preprint arXiv:2510.06188*, 2025.

A. Howard, A. I. Humayun, A. Chow, R. Holbrook, Sushmit, and Tahsin, "Bengali.AI Speech Recognition," *Kaggle*, 2023. [Online]. Available: `https://kaggle.com/competitions/bengaliai-speech`

M. H. Nayem et al., "An overview of Bengali speech recognition," *ResearchGate*, 2023.

M. A. Rahman et al., "Phonological variation and linguistic diversity in Bangladeshi dialects," *Frontiers in Language Studies*, 2024.

A. Shawon et al., "A Regional Corpus of Bengali Spontaneous Speech Across Dialects," *arXiv preprint arXiv:2510.24096*, 2025.

S. Sultana, M. S. Rahman, and M. Z. Iqbal, "Recent advancement in speech recognition for Bangla: A survey," *IJACSA*, vol. 12, no. 3, 2021.

World Population Review, "Bangladesh Population 2025," 2025.