Regional Bengali Dialect to Standard ASR: An Ensemble Approach with Fine-Tuned Whisper Models

Ruwad Naswan[1] **Shadab Tanjeed Ahmad**[2] **Abrar Zahin Raihan**[1]
Department of Computer Science & Engineering
[1]Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[2]Islamic University of Technology, Dhaka, Bangladesh
{ruwad45678, zaheenraian}@gmail.com shadabtanjeed@iut-dhaka.edu

Abstract

This paper presents a method for transcribing 20 regional Bangladeshi dialects into standard Bangla text. We fine-tune OpenAI's Whisper models (`bengaliAI/whisper-medium` and `whisper-medium-regional`) on 3,350 dialectal audio recordings and explore multiple fine-tuning strategies, including Frozen Decoder, Full Fine-Tuning, Dialect Classifier, and Regional Fine-Tuning. The approach incorporates standard audio preprocessing and selective data augmentation for under-represented dialects. Our experiments demonstrate that incorporating dialect-specific embeddings significantly improves transcription accuracy, particularly for low-resource regional varieties. The results highlight the effectiveness of pre-trained multilingual models when combined with targeted fine-tuning and region-aware modeling for robust Bangla ASR.

1 Introduction

Bangladesh exhibits remarkable linguistic diversity, with over 20 regional dialects characterized by distinct phonetic features, vocabulary, and pronunciation patterns. While standard Bangla serves as the formal written language, regional dialects dominate everyday communication, creating a substantial gap in speech technology accessibility.

Automatic Speech Recognition (ASR) systems trained primarily on standard Bangla often fail to handle these dialectal variations, limiting their practical utility across Bangladesh's diverse linguistic landscape. The challenge extends beyond simple accent differences to encompass vocabulary variation, phonetic realization, and grammatical structures.

This work addresses the problem of developing robust ASR models capable of transcribing dialectal Bangla speech into standard written Bangla. Our approach leverages pre-trained Whisper models (`bengaliAI/whisper-medium` and `whisper-medium-regional`) and explores multiple fine-tuning strategies, including Frozen Decoder, Full Fine-Tuning, Dialect Classifier, and Regional Fine-Tuning. Additionally, we apply standard audio preprocessing and selective data augmentation for under-represented dialects.

The dataset used in this study comprises 3,800 audio recordings, including 3,350 training samples and 450 test samples, spanning diverse regions such as Chittagong, Barisal, Sylhet, and Mymensingh. Regional sample sizes range from 21 to 401 recordings, reflecting the natural imbalance across dialects.

Our contributions include a systematic fine-tuning methodology for Whisper models on multi-dialectal Bangla speech and a dialect-aware classifier for improving transcription accuracy across regional variations.

2 Related Work

2.1 Dialectal Speech Recognition

Dialectal ASR has gained increasing attention as systems expand beyond standard languages. For Bengali, several recent works address regional standardization: Samin et al. introduced BanglaDialecto, an end-to-end framework for dialect conversion; Hassan et al. released RegSpeech12, a large spontaneous speech corpus covering multiple dialects; and Biswas et al. proposed a unified denoising–adaptation framework for self-supervised Bengali dialectal ASR, emphasizing robust preprocessing for noisy data. Beyond Bengali, Talafha et al. benchmarked Whisper on diverse Arabic dialects, highlighting both strengths and limitations of foundation models. Collectively, these studies show that while multilingual models provide strong baselines, targeted adaptation remains essential for phonetic and lexical variability.

2.2 Dialect Translation and Standardization

Parallel work focuses on text-based dialect translation. Sultana et al. introduced ONUBAD, a dataset for converting regional dialects to standard Bangla. Faria et al. developed Vashantor, a large multilingual benchmark addressing semantic preservation in dialect translation. Khandaker et al. explored bidirectional standard–dialect translation, while Dip et al. focused on the Rangpur dialect. Paul et al. improved dialect detection using BERT, LLMs, and explainable AI, highlighting its importance as a precursor to effective standardization. Dipto et al. evaluated whether ASR foundation models sufficiently capture dialectal features, finding that specialized adaptation remains necessary.

2.3 Foundation Models for ASR

Large-scale foundation models have reshaped ASR. Radford et al. introduced Whisper, trained on 680,000 hours of multilingual data, enabling strong zero-shot performance. Baevski et al. proposed wav2vec 2.0, a self-supervised framework foundational for low-resource ASR. Integration with LLMs is increasingly explored: Mittal et al. presented SALSA, a synchronous ASR–LLM aggregation method; Chen et al. proposed HyPoradise, an open baseline using LLMs for correction; Radhakrishnan et al. developed Whispering LLaMA for cross-modal error correction; and Wu et al. investigated decoder-only speech-to-text architectures. Building on this, our work combines complementary fine-tuning strategies—frozen decoder, full fine-tuning, and regional adaptation—with ROVER-based ensemble aggregation to achieve improved Bengali dialectal ASR performance.

3 Methodology

Our approach combines three key components: (1) Implementing data augmentation and preprocessing strategies (2) Fine tuning pre-trained OpenAI Whisper Model and (3) Ensembling the best performing models

3.1 Model Architecture

We employ the Whisper architecture, an encoder-decoder transformer model pre-trained on 680,000 hours of multilingual and multitask supervised data. The architecture consists of an audio encoder that processes log-Mel spectrogram inputs and a text decoder that generates transcriptions autoregressively.

For this study, we utilize pre-trained models, specifically bengaliai-regional-asr_whisper-medium and bengaliai-asr_whisper-medium, each comprising approximately 800 million parameters. These models already provide a solid ground for our task of converting Bangla regional speech into standard text, enabling effective fine-tuning for regional-to-standard ASR.

3.2 Data Collection and Preprocessing

The dataset contains 3,350 training and 450 test recordings across 20 regional Bangla dialects. Each audio sample is paired with a standard Bangla transcription. All recordings are 16 kHz mono WAV files totaling 3.90 hours of speech, with an average duration of 4.2 seconds. The vocabulary includes 590 unique words, with transcripts averaging 5.4 words (29.7 characters) and a punctuation ratio of 3.4%. Duration exhibits a long-tail distribution (skewness = 4.44), with 90% of utterances between 3–6 seconds, reflecting the brief conversational nature of regional speech.

The dataset is notably imbalanced: regions such as Comilla, Rajshahi, Noakhali, Feni, and Barishal have fewer than 100 samples (Figure 1). To address this, we incorporated the RegSpeech12 corpus, which provides dialectal audio with dialect-level transcriptions for 12 regions. As these transcriptions are in local dialects, we converted them to standard Bangla using the Gemini 2.5 Flash API. For six regions, we added 30–70 additional samples to improve coverage and reduce bias toward majority dialects.

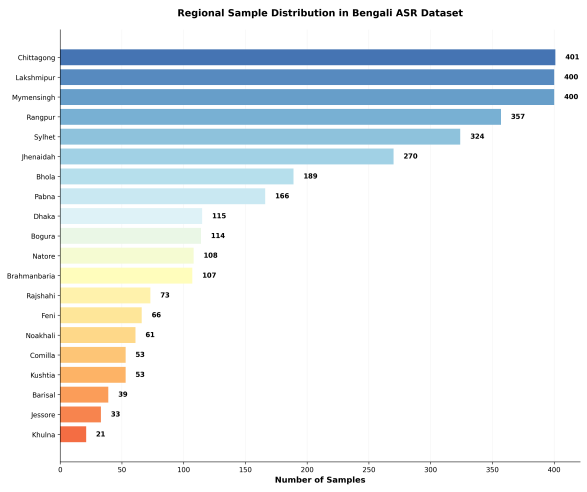

Figure 1: Regional distribution of training audio files

For preprocessing, all recordings were standardized to 16 kHz mono. Denoising used dynamic

spectral gating via `librosa`, followed by amplitude normalization to $-3\,dB$. Given a signal $x(t)$ with RMS amplitude A_{rms}, the scaling factor is

$$s = \frac{10^{(-3/20)}}{A_{\text{rms}}},$$

yielding the normalized signal $x_{\text{norm}}(t) = s \cdot x(t)$. Recordings shorter than 10 seconds were padded with 3.5 seconds of silence, and extremely short clips (<1 s) were removed to avoid training instability.

3.3 Data Augmentation

To mitigate dataset imbalance and improve robustness, we applied audio augmentations during training, including time stretching, pitch shifting, noise injection, and volume adjustment. Each was applied with probability $p = 0.3$, providing variability while preserving audio quality. All transformations maintain the phonetic content to avoid semantic distortion. Augmentation was applied only during training, leaving the evaluation distribution unchanged.

3.4 Fine-tuning Strategy

For our fine-tuning strategy, we explored four distinct approaches applied to both the `bengaliAI/whisper-medium` ASR model and the `bengaliAI/whisper-regional-medium` variant: (1) Frozen Decoder, (2) Full Fine-Tuning, (3) Dialect Classifier, and (4) Regional Fine-Tuning.

Frozen Decoder

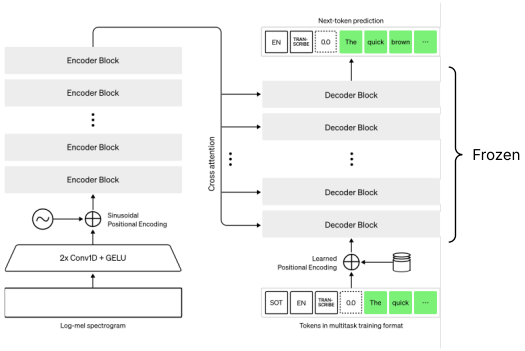

Figure 2: Whisper model architecture with frozen decoder.

The frozen decoder approach leverages transfer learning by fine-tuning only the encoder while keeping the decoder fixed. We apply this to the BengaliAI Whisper Medium model (Tugusgui et al., INTERSPEECH 2023), pre-trained on 1,177 hours of standard Bengali audio–text pairs. This pre-training enables the decoder to reliably convert acoustic embeddings into standard Bengali text.

Freezing the decoder preserves its learned linguistic knowledge—Bengali grammar, vocabulary, and orthography—while adapting only the encoder to capture regional acoustic variations. This is effective because dialectal Bengali shares most linguistic structure with standard Bengali; differences are primarily in pronunciation and phonetic patterns handled by the encoder. The frozen decoder ensures consistent standard output while the encoder maps diverse dialectal features appropriately.

This approach also reduces trainable parameters compared to full fine-tuning, lowering overfitting risk on the limited dialectal dataset while retaining strong text generation. In our ensemble, the frozen decoder contributed complementary strengths by prioritizing linguistic structure preservation over aggressive acoustic adaptation.

Full Fine-Tuning

Full fine-tuning updates all model parameters—encoder and decoder—on the competition dataset, enabling adaptation to regional dialects. Using the BengaliAI Whisper Medium model pretrained on standard Bengali audio–text, both encoder and decoder jointly adjust to dialectal speech.

This approach leverages the full augmented dataset, including RegSpeech12 converted via Gemini 2.5 Flash, with augmentations (time-stretch, pitch-shift, noise, volume adjustment, random cropping) and weighted sampling to mitigate dialect imbalance. Updating all parameters handles acoustic variations and subtle text differences for accurate transcription.

Full fine-tuning provides maximum task-specific flexibility, though regularization is needed to avoid overfitting. It delivers robust baseline transcription to the ensemble, excelling with comprehensive parameter adaptation.

Dialect Classifier

To incorporate dialect-specific information into the Whisper encoder, we introduce a regional embedding module on top of the `bengaliAI/whisper-medium-regional` model (Figure 4). Each dialect region is assigned an integer label (0–20), mapped to a 64-dimensional embedding vector. Using a 64-dimensional adapter rather than the full 1024-dimensional encoder space provides a lightweight representation:

the embedding matrix requires only 64×20 parameters, compared to 1024×20.

Because the Whisper encoder outputs hidden states of dimension $d_{model} = 1024$, a linear projection layer maps the 64-dimensional embedding to the encoder dimension. The projected vector is injected into the encoder via additive modulation:

$$h' = h + W_{proj}\, e_r,$$

where h is the encoder hidden state, e_r the regional embedding, and W_{proj} the projection matrix. As the projection is randomly initialized, its output may initially have high variance, so a normalization layer is applied afterward for stability.

For region classification, the augmented encoder outputs are mean-pooled across time to produce an utterance-level embedding, which is passed through a linear classifier producing logits for the 20 regions. Alternatives such as MLPs or dropout can also be used to increase robustness.

Training follows a multi-task objective combining ASR and regional classification losses:

$$\mathcal{L}_{total} = \mathcal{L}_{ASR} + \alpha\, \mathcal{L}_{region},$$

with $\alpha = 0.3$. The ASR loss penalizes transcription errors, while regional classification uses cross-entropy. The model is optimized with a learning rate of 1×10^{-5}, weight decay 0.01, label smoothing 0.1, for 8 epochs.

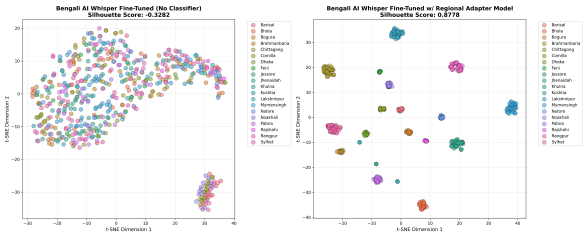

Figure 3: Regional Classifier Adapter showing t-SNE embeddings before and after adaptation

Regional Fine-Tuning

For the regional fine-tuning experiments, we utilized the `bengaliAI/whisper-medium-regional` model. Standard audio preprocessing was applied to all recordings, but no additional data augmentation techniques were used. Fine-tuning was performed exclusively on the provided competition dataset. The model was trained for 5 epochs with a learning rate of 1×10^{-5} and a weight decay of 0.01.

3.5 Ensembling Methods

To maximize transcription accuracy and leverage complementary model strengths, we employed the Recognizer Output Voting Error Reduction (ROVER) ensemble strategy, a consensus-based technique that selects predictions through weighted voting based on pairwise similarity scores.

3.5.1 Ensemble Components

Our final ensemble consists of four Whisper Medium variants: frozen decoder fine-tuning (encoder-only adaptation), full fine-tuning on standard Bangla (complete parameter updates), full fine-tuning on regional Bangla (specialized dialectal knowledge from Ben10 pre-training), and regional classifier adapter (multi-task learning with encoder-level regional conditioning). All models were trained on the augmented competition dataset with weighted sampling and RegSpeech12 external data.

3.5.2 ROVER Algorithm

The ROVER algorithm collects predictions from all four models, computes pairwise Levenshtein similarity scores, and selects the prediction maximizing average similarity to all candidates with learned weights applied based on validation performance. This consensus-based approach effectively reduces individual model errors by filtering outliers while preserving correct predictions.

3.5.3 Ensemble Performance

The ROVER ensemble significantly outperformed individual variants, achieving 0.93509 NLS on the public leaderboard and 0.91782 NLS on the private leaderboard. Each model captures different aspects of dialectal speech recognition—frozen decoder preserves linguistic structure, standard fine-tuning provides robust baselines, regional fine-tuning handles dialectal nuances, and classifier adapters offer regional awareness. This diversity enables ROVER to produce more reliable transcriptions with strong generalization across unseen dialectal samples.

3.6 Post-processing with LLMs

Raw ASR outputs often contain errors due to phoneme confusion, incomplete dialect normalization, and out-of-vocabulary words. We experimented with LLM-based post-processing to refine transcriptions and standardize dialectal Bangla to formal Bangla using prompts like: *"The following is a Bangla speech transcription that may con-*

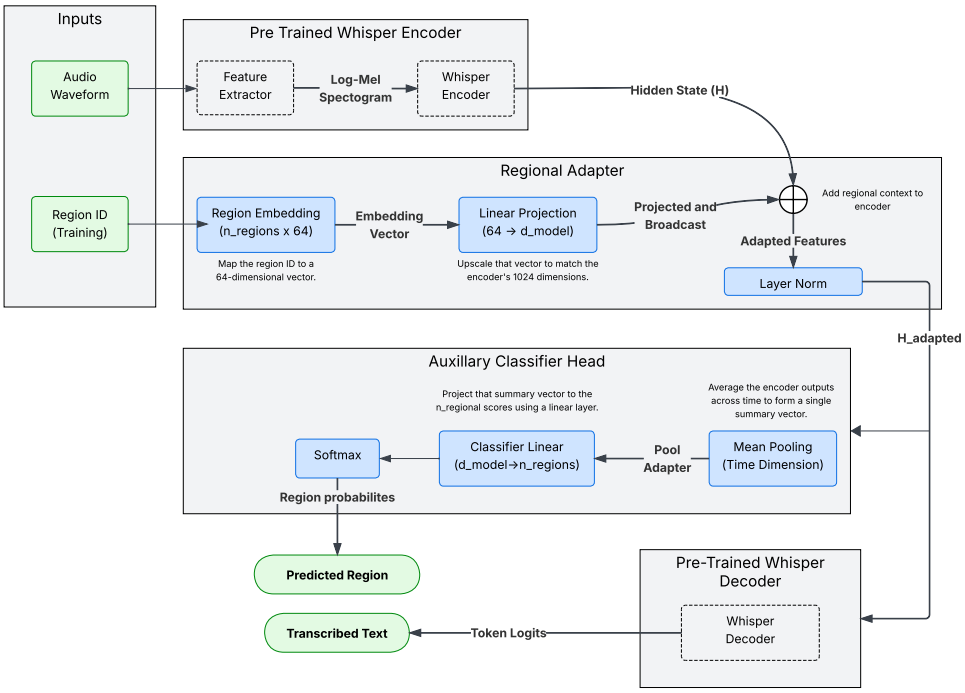

Figure 4: Classifier adapter architecture used in the regional Whisper model.

tain errors or dialectal expressions. Please correct any transcription mistakes and convert dialectal Bangla to standard formal Bangla while preserving the original meaning: [TRANSCRIPTION]"

However, LLM post-processing proved unreliable for production due to non-deterministic outputs, lack of acoustic context leading to plausible but incorrect corrections, and high computational cost. Consequently, our final ensemble relies on the ROVER algorithm without LLMs, achieving strong performance through diverse model variants and weighted voting.

3.7 Inference Pipeline

The inference pipeline operates in four stages: (1) audio preprocessing including resampling to 16 kHz mono, denoising via dynamic spectral gating, and -3 dB normalization, (2) Log-Mel spectrogram extraction for acoustic feature representation, (3) parallel inference across all four ensemble model variants, and (4) ROVER-based ensemble aggregation through weighted voting followed by text normalization to ensure proper UTF-8 encoding and standard Bangla orthography. The complete pipeline achieves approximately 0.7× real-time speed on GPU, enabling efficient batch processing of dialectal speech while maintaining high transcription accuracy across all 20 regional variants.

4 Results and Analysis

4.1 Experimental Setup

All experiments are conducted using the Hugging Face Transformers library and PyTorch framework. Training is performed on NVIDIA P100 GPU with mixed-precision (fp16) training for computational efficiency.

4.2 Evaluation Metric

Model performance is measured using Normalized Levenshtein Similarity (NLS):

$$NLS(r, p) = 1 - \frac{LevenshteinDistance(r, p)}{\max(|r|, |p|)}$$

where r represents the reference transcription and p the predicted transcription. The final score is the mean NLS across all test samples.

4.3 Results

Table 1 shows the different scores achieved through our methodology.

Evaluation of fine-tuning strategies shows notable performance differences across private and public datasets. The Dialect Classifier achieves the highest scores (Private NLS: 0.8996, Public NLS:

Approach	Private NLS Score	Public NLS Score
Frozen Decoder	0.89496	0.91996
Full Fine-Tuning	0.89294	0.90581
Dialect Classifier	0.89959	0.91488
Regional Fine-Tuning	0.88103	0.91096

Table 1: NLS scores for different fine-tuning strategies on private and public sets.

0.9149), indicating that explicitly incorporating regional information strengthens the model's ability to capture dialectal variations. Frozen Decoder and Full Fine-Tuning are competitive, with the former slightly better on private data and the latter better on public data. Regional Fine-Tuning, while helpful for specific dialects, has slightly lower private scores due to limited training data. Overall, augmenting Whisper with dialect-aware components substantially improves transcription quality, emphasizing the value of region-specific modeling in Bangla ASR.

4.4 Competition Performance

Our final submission achieves competitive performance on the AI-FICATION hackathon:

Leaderboard	NLS Score
Public Leaderboard (30%)	0.93518
Private Leaderboard (70%)	0.91782
Final Ranking (Private LB)	2nd Position

Table 2: Competition performance on public and private test sets

The slight decrease from public to private leaderboard indicates robust generalization with minimal overfitting. Our approach successfully handles the diverse dialectal variations present in the test set, demonstrating the effectiveness of combining fine-tuned Whisper models with LLM-based post-processing for dialectal Bangla ASR.

5 Error Analysis

Our error analysis reveals that despite achieving over 90% perfect transcriptions (zero WER), the system faces specific challenges primarily related to out-of-vocabulary (OOV) dialectal words and the inherent complexity of regional speech variations.

5.1 Model Performance Overview

The ensemble model demonstrates exceptional accuracy with the majority of predictions achieving perfect transcription. Analysis of the embedding space through t-SNE visualization shows that errors are scattered randomly without clustering patterns, indicating no systematic failure modes or acoustic confusion regions. This suggests the model has learned robust, generalizable representations rather than memorizing training patterns.

However, isolated errors do occur, primarily in two areas:

5.2 Out-of-Vocabulary Dialectal Terms

A major challenge arises from rare dialect-specific words that are:

- **Under-represented:** Many region-specific terms appear too infrequently for the model to learn their acoustic–phonetic patterns reliably.
- **Orthographically Variable:** Dialectal words often lack standardized spellings, leading to mishearings or plausible but incorrect transcriptions.
- **Phonetically Ambiguous:** Some dialectal terms closely resemble common standard Bangla words, causing the model to default to higher-frequency alternatives.

5.3 Edge Cases in Utterance Length

Performance analysis across word counts reveals that while the model excels at standard-length utterances (5-7 words with >0.8 similarity), isolated failures occur at edge cases:

- **Very Short Utterances:** Sentences with 3-4 words occasionally lack sufficient context for the model to disambiguate between similar-sounding words.
- **Specific Acoustic Conditions:** Some short utterances with background noise or rapid speech exhibit reduced accuracy, though these represent a small fraction of errors.

5.4 Generalization and Robustness

The small performance gap between the public (0.93509 NLS) and private (0.91782 NLS) leaderboards indicates strong generalization with minimal overfitting. The slight drop is likely due to:

- Variation in test set difficulty,

- More challenging dialectal samples in the private set,
- Edge cases involving OOV terms and rare acoustic conditions.

The absence of consistent error clustering further shows that the ensemble effectively reduces individual model biases and remains robust across dialects. Future gains will require broader dialectal coverage and targeted collection of rare vocabulary to mitigate OOV-related errors.

6 Conclusion

We present a comprehensive approach for multi-dialectal Bangla ASR using the bengaliai-regional-asr_whisper-medium and bengaliai-asr_whisper-medium models with targeted fine-tuning strategies. Four methods—Frozen Decoder, Full Fine-Tuning, Dialect Classifier, and Regional Fine-Tuning—are explored alongside standard preprocessing and selective augmentation for under-represented regions.

Our results show that dialect-specific embeddings via the Dialect Classifier provide the strongest gains, underscoring the value of region-aware modeling. Regional Fine-Tuning further benefits low-resource dialects, while full adaptation offers solid generalization. Overall, pre-trained Whisper models, when carefully adapted and augmented with dialectal knowledge, can effectively transcribe regional Bangla speech into standard form.

Despite strong performance, several limitations remain. Regional coverage is constrained by dataset imbalance; Whisper-based training is computationally intensive; and our multi-task setup does not incorporate more advanced ASR–LLM fusion techniques.

Future work includes integrating external language models for normalization (e.g., KenLM, Vicuna-series LLMs), exploring deeper ASR–LLM fusion such as SKIP-SALSA, expanding dialectal speech coverage, and incorporating speaker attributes like gender to improve robustness and fairness.

Acknowledgments

We thank the AI-FICATION organizing committee and the Department of Electronics & Telecommunication Engineering, CUET, for organizing this competition.