Fine-Tuning Bengali Speech Recognition for Regional Dialect Diversity

Md. Sajjad Hossain, Ashraful Islam Paran and Symom Hossain Shohan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904031,u1904029,u1904048}@student.cuet.ac.bd

***Abstract*—The use of automatic speech recognition (ASR) systems in human-computer interaction is vital, but because these systems are trained on standard language datasets, they frequently have difficulty distinguishing regional dialects. We tackle this problem in our work by concentrating on Bengali Speech Recognition for Regional Dialects. We adjust the OpenAI Whisper small model by fine-tuning its hyperparameters. Our finetuned model achieved 5th rank with private score of 0.75438.**

I. Methodology

[1]In this task, the dataset provided contains audio file and the correspondig transcript. The audio file and transcript were preprocessed independently in this experiment. Subsequently, the whisper small model is fine-tuned with the appropriate hyperparameters using the preprocessed data. Figure 1 shows the general workflow for this approach.

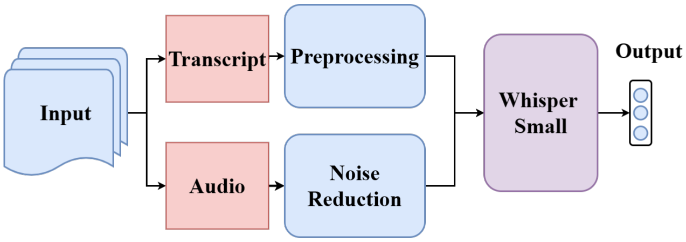


Fig. 1. Schematic process of Bengali Speech Recognition for Regional Dialect

A. Data Preprocessing

In the data preprocessing phase, a number of steps were followed for cleaning and normalizing the transcripts as well as the audio files. Initially, all punctuation, emojis, and extra spaces were eliminated from the transcripts. After that, the csebuetnlp/normalizer [2] was used to normalize each and every transcript. To improve the quality of the training data, short transcripts were removed from the dataset. The audio files were subjected to noise reduction in order to improve the audio quality overall. These preprocessing steps were essential to preparing the data for the model's training and evaluation.

B. Hyperparameter Tuning

In this task, the Whisper small [3] speech recognition model by `OpenAI`[1] was fine-tuned and employed for our particular use case. In the course of fine-tuning, we set the evaluation strategy to 'epoch', the learning rate to 5e-5, the weight decay to 1e-2, and the number of epochs to 2. To achieve optimal performance on the validation set, we tuned and empirically experimented before selecting these hyperparameters. The labelled dataset—which was comprised of audio samples and the transcriptions that matched them—was used to train the model during this procedure.

TABLE I
HYPERPARAMETER SETTINGS FOR THE EMPLOYED MODEL

Hyperparameter	Value
Epoch	2
Learning Rate	5e-5
Weight Decay	1e-2
Evaluation Strategy	epoch

II. Results and Analysis

The models are evaluated based on the mean Word Error Rate [4] which is the ratio of the sum of all three types of errors (substitutions, deletions, insertions) to the total number of words in the reference transcript. WER is calculated using the formula:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript.

Table II exhibits the evaluation results of various models that we submitted to the competition [5]. The results demonstrate that the finetuned model outperforms the base models whisper-small and whisper-medium by a wide margin. The public score for whisper-small is 0.94180, and the private score is 0.91583. Whisper-medium performs somewhat worse, with a public score of 0.99938 and a private score of 0.99467.Our whisper-small model, fine-tuned for regional Bengali dialects, achieves a public score of 0.76412 and a private score of 75438, respectively.

Figure 2 shows the loss value on the training set. The loss curve begins at a high magnitude and steadily declines with each epoch. The curve rises after it converges.

[1] https://www.openai.com/

TABLE II
PERFORMANCE OF THE EMPLOYED MODELS ON THE TEST SET

Model	Public Score	Private Score
Whisper-small	0.94180	0.91583
Whisper-medium	0.99938	0.99467
Whisper-small-finetuned	0.76412	0.75438

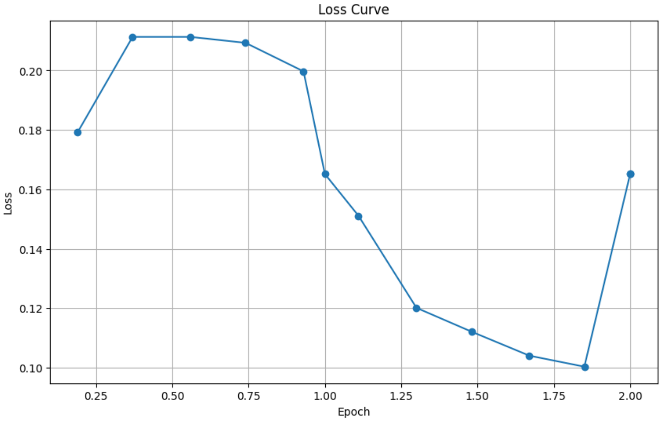
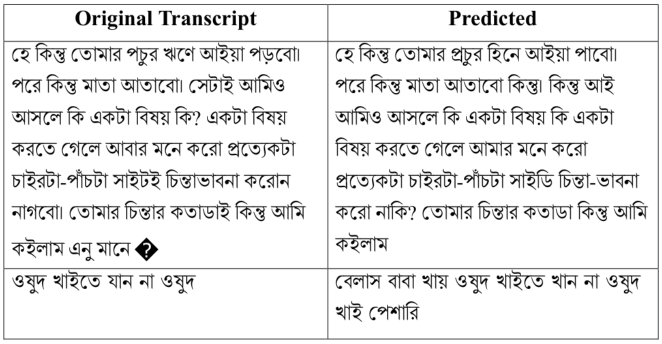


Fig. 2. Loss Curve on training

Figure 3 depicts some examples of original transcript and predicted transcripts of the model.



Original Transcript	Predicted
হে কিন্তু তোমার পচুর ঋণে আইয়া পড়বো। পরে কিন্তু মাতা আতাবো। সেটাই আমিও আসলে কি একটা বিষয় কি? একটা বিষয় করতে গেলে আবার মনে করো প্রত্যেকটা চাইরটা-পাঁচটা সাইটই চিন্তাভাবনা করোন নাগবো। তোমার চিন্তার কতাডাই কিন্তু আমি কইলাম এনু মানে �	হে কিন্তু তোমার প্রচুর হিনে আইয়া পাবো। পরে কিন্তু মাতা আতাবো কিন্তু। কিন্তু আই আমিও আসলে কি একটা বিষয় কি একটা বিষয় করতে গেলে আমার মনে করো প্রত্যেকটা চাইরটা-পাঁচটা সাইডি চিন্তা-ভাবনা করো নাকি? তোমার চিন্তার কতাডা কিন্তু আমি কইলাম
ওষুদ খাইতে যান না ওষুদ	বেলাস বাবা খায় ওষুদ খাইতে খান না ওষুদ খাই পেশারি

Fig. 3. Sample prediction with original transcript

Pre-trained models struggle to effectively recognize Bangla local dialects since they are typically trained on pure Bangla datasets rather than local dialects. This study closes the gap in automatic voice identification for local dialects.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] T. Hasan, A. Bhattacharjee, K. Samin, M. Hasan, M. Basak, M. S. Rahman, and R. Shahriyar, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 2612–2623, Association for Computational Linguistics, Nov. 2020.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[4] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 20–24, 2018.

[5] R. S. F. S. M. J. I. S. T. Md. Rezuwan Hassan, Mohaymen Ul Anam, ": Asr for regional dialects," 2024.