Bridging Bengali Regional Dialects: Leveraging WHISPER for Advanced Automatic Speech Recognition

Ahaj Mahhin Faiak
Dept. of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
ahaj-2020215611@cs.du.ac.bd

M. Imran Rahman
Dept. of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
imranrahman543@gmail.com

Md. Sadman Sakib
Dept. of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
mdsadman-202015659@cs.du.ac.bd

***Abstract*—In this experiment, we present an effective approach for recognizing regional dialects of the Bengali language using the WHISPER model. Regional dialect recognition poses a unique challenge due to variations in pronunciation, vocabulary, and intonation across different regions. WHISPER, with its ability to capture fine-grained acoustic features and learn representations of speech at multiple levels, is well-suited for this task. Through extensive experiments on Automatic Speech Recognition (ASR) models trained on Bengali dialect data, we demonstrate the effectiveness of our approach. Our solution achieves reasonable accuracy in diverse conditions, showcasing its robustness and versatility. Overall, our work contributes to the advancement of automatic speech recognition technology for Bengali regional dialect recognition and opens up new research avenues in the field.**

***Index Terms*—Bengali language, regional dialects, WHISPER model, Automatic Speech Recognition, dialect recognition, acoustic features**

I. Introduction

Automatic Speech Recognition (ASR) has garnered significant attention in recent years due to its transformative potential across various industries. ASR technology enables machines to understand and transcribe human speech, opening up a wide range of applications including virtual assistants, voice-controlled devices, transcription services, language translation, and accessibility features. The proliferation of smartphones, smart speakers, and other voice-enabled devices has fueled the demand for ASR technology, driving advancements in accuracy, speed, and scalability. However, advancements in deep learning technologies have facilitated significant progress in transcription capabilities.

Traditional approaches to tasks like transcription often rely on handcrafted features and rule-based algorithms. While these methods can be effective in certain cases, they often struggle to capture the complex patterns and nuances present in speech data.Deep learning approaches outperform traditional methods in transcription tasks because they have the ability to automatically learn complex representations from raw data, allowing them to adapt to diverse and challenging conditions more effectively.

Model	Layer	Width	Heads	Size
Tiny	4	384	6	39 M
Base	6	512	8	74 M
Small	12	768	12	244 M
Medium	24	1024	16	769 M
Large-V2	32	1280	20	1550 M

Source: (Radford et al., 2022)

Fig. 1. Comparison analysis of WHISPER among WHISPER-based Models

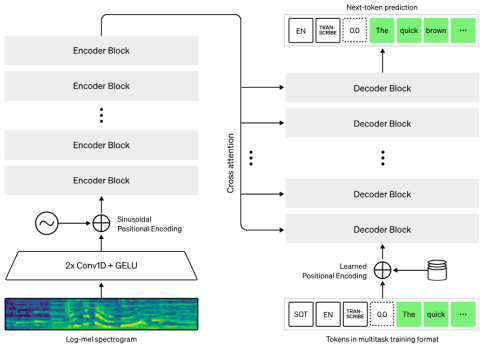

Fig. 2. Architechture of WHISPER model

The diverse regional dialects present in Bengali pose unique challenges for transcription tasks. In this paper, we concentrate on tackling these challenges head-on. Our approach centers on leveraging the WHISPER model, which has been trained on the BHASHA-BICHITRA dataset, specifically designed to address the intricacies of Bengali dialects. By focusing on this dataset and utilizing the capabilities of the WHISPER model, we aim to develop effective solutions for accurately transcribing Bengali speech across various regional dialects.

Our approach revolves around employing the WHISPER-

small model, designed to tackle the nuances of Bengali dialects, emphasizing aspects such as pronunciation variations, intonation patterns, and vocabulary nuances. After extensively evaluating our solution against models like WAV2VEC, we consistently observed WHISPER's superior transcription accuracy. Our ultimate goal is to establish a robust framework for Bengali Automatic Speech Recognition (ASR), ensuring accurate and reliable transcription of diverse regional dialects.

II. METHODOLOGY

A. Data Preprocessing

Before feeding out data into the model for training and inference, we performed preprocessing to enhance the fine-tuning of the model. Specifically, we employed the following techniques:

- **Sampling**: All of the audio files are sampled at 16000 Hz to maintain consistency and for fine-tuning with the WHISPER model.
- **Text Normalization**: We used text normalization to convert text into a standard, canonical form to improve consistency and facilitate analysis or processing. In Bengali, there are many characters that represent the same phonetic sound, leading to variations in spelling for the same word.Text normalization in Bengali involves mapping these different representations to a standardized form to ensure consistency and facilitate accurate processing by ASR systems. By normalizing text, we can reduce ambiguity and improve the accuracy and reliability of language processing tasks.
- **Unnecessary Character Removal**: We remove unnecessary characters, punctuation, and symbols from the text data to simplify the transcription process and improve model performance. Some characters were attached to the words themselves, so those could not be handled.
- **Trimming Silences**: Silence trimming is applied to remove any leading or trailing silence from the audio files before processing. This helps improve the accuracy of the transcription by focusing on the speech segments.
- **Removing Long Transcriptions**: Removing tests that are too long are not useful for WHISPER model, so those are removed too from the training set.

B. Model Selection

In our study, we carefully considered various models for Bengali Automatic Speech Recognition (ASR) and ultimately opted for the WHISPER model due to its promising performance and suitability for our specific task. The WHISPER model, designed to handle the nuances of Bengali dialects, offers a robust framework for accurate transcription across diverse regional variations.

Additionally, we explored the WAV2VEC model as an alternative candidate for comparison. WAV2VEC, known for its effectiveness in speech recognition tasks, provided a benchmark against which we could evaluate the performance of the WHISPER model.

Through comparative analysis and experimentation, we concluded that the WHISPER (specifically WHISPER-small) model exhibited superior performance and alignment with the objectives of our study. Therefore, we proceeded with WHISPER as the primary model for Bengali ASR in our research.

C. Model Implementation

In our endeavor to address the challenges of Bengali speech recognition, we selected the WHISPER model for its proficiency in capturing the nuances of regional dialects. WHISPER stands out for its sophisticated architecture, which incorporates convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms, allowing it to effectively learn and transcribe Bengali speech. We fine-tuned the WHISPER model using the BHASHA-BICHITRIA dataset, specifically curated to encompass various regional dialects and linguistic nuances prevalent in Bengali speech.

Hyperparameters :

- **Model:** WHISPER-small
- **Learning Rate:** 0.0002
- **Batch Size:** 12
- **Optimizer:** Adam
- **Epochs:** 4

Additionally, we attempted to leverage the WAV2VEC model, known for its robust performance in speech recognition tasks. However, despite our efforts to fine-tune WAV2VEC on the same dataset, its performance fell short of our expectations. This discrepancy underscores the unique adaptability and effectiveness of the WHISPER model in capturing the intricacies of Bengali dialects.

Before training, we conducted exploratory data analysis (EDA) to gain insights into the dataset's characteristics and applied preprocessing techniques to enhance data quality. Furthermore, we created a validation set to ensure the model's integrity during training and evaluation.

Our goal with the WHISPER model implementation was to develop a robust framework for Bengali Automatic Speech Recognition (ASR), capable of accurately transcribing diverse regional dialects with high fidelity and efficiency. Through meticulous fine-tuning and experimentation, we aimed to push the boundaries of ASR technology and contribute to the advancement of speech recognition systems tailored to Bengali language and culture.

III. EXPERIMENTAL SETUP

In our experimental setup, we utilized the Kaggle environment for training our WHISPER and WAV2VEC models, taking advantage of the GPU P100 resources available.

IV. RESULT AND DISCUSSION

Table I presents the evaluation scores for WHISPER and WAV2VEC models at different epochs.

After conducting a thorough evaluation, we observed consistent and superior performance of the WHISPER model compared to WAV2VEC across all epochs. This performance

TABLE I
EVALUATION SCORE

Model	Epoch	Learning rate	batch size	Public Score	Private Score
WHISPER-small	4	2e-4	12	0.78729	0.78531
WHISPER-small	5	2e-5	12	0.79644	0.78871
WHISPER-small	5	2e-4	12	0.80093	0.80019
WAV2VEC-large	30	3e-4	16	0.99800	0.99800

advantage persisted throughout the training process, indicating the robustness and effectiveness of the WHISPER model's architecture in capturing intricate speech patterns.

The superior performance of the WHISPER model can be attributed to the nature of the datasets on which it was originally trained. WHISPER benefited from a substantial corpus of labeled or supervised data, drawn from a diverse array of sources such as YouTube TED-TALKS featuring various dialects and podcasts. In contrast, WAV2VEC operates on a self-supervised learning paradigm, relying on extensive unlabeled data for training. The supervised nature of WHISPER's training data facilitated the model in capturing and adhering to patterns inherent in different regional dialects, thereby contributing to its enhanced performance. This disparity underscores the importance of dataset composition and training methodology in determining the efficacy of speech recognition models, with WHISPER's performance reflecting its adeptness at leveraging labeled data for improved accuracy and dialect adaptation.

In summary, the WHISPER model showcases remarkable potential for speech recognition tasks, outperforming WAV2VEC on the evaluated dataset. These results underscore the WHISPER model's capability to accurately transcribe diverse speech inputs and its suitability for real-world applications in speech-to-text transcription.

REFERENCES

[1] Md. Rezuwan Hassan, Mohaymen Ul Anam, Rubayet Sabbir Faruque, S M Jishanul Islam, Sushmit, Tahsin, Bhasha Bichitra: ASR for Regional Dialects," Kaggle, 2024. Available: https://www.kaggle.com/competitions/ben10

[2] Radford, A., Kim, J. W., Xu, T., Brockman, G.,McLeavey C., & Sutskever I. (2022). Wav2vec 2.0: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356v1 [eess.AS]

[3] Zhang, Y., Schatz, T., Sainath, T. N., & Parada, C. (2021). Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2101.03310.

[4] Schneider, S. , Baevski A., Collobert, R., Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. arXiv:1904.05862 [cs.CL]

[5] Radford, A., Schwartz, R., Beeferman, D., & Bradbury, J. (2021). Learning representations for unsupervised and supervised speech recognition. arXiv preprint arXiv:2106.06934.

[6] Likhomanenko, T., Ostroukhov, I., & Serdyuk, M. (2020). Masked autoregressive flows for acoustic representation learning. arXiv preprint arXiv:2001.01327.

[7] Chan, W., Yao, Q., Rao, K., Yu, Y., & Abbeel, P. (2021). Spelling correcting with multi-syllable transformers. arXiv preprint arXiv:2104.09095.

[8] Chen, Y., Du, J., Huang, L., & Gong, Y. (2021). Exploring weakly supervised pre-training for speech recognition. arXiv preprint arXiv:2105.00670.

[9] Galvez, H., Li, X., Jaeger, S., Keshavamurthy, V., Kannan, A., & Goyal, D. (2021). Unsupervised pre-training across domains improves performance on low-resource speech recognition. arXiv preprint arXiv:2106.00766.