7th Bangladesh Conference on Industrial Engineering and Operations Management
December 21-22, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/BA07.20240161
Published: December 21, 2024

Enhancing Bangla Local Speech-to-Text Conversion Using Fine-Tuning Wav2vec 2.0 with OpenSLR and Self-Compiled Datasets Through Transfer Learning

Sk Muktadir Hossain, Md Rahat Rihan
Ahmed Imtiaz
Department of Computer Science and Engineering
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
21-44989-2@student.aiub.edu, 21-45033-2@student.aiub.edu
21-44914-2@student.aiub.edu

Pritam Khan Boni, Dipta Justin Gomes
Lecturer, Department of Computer Science
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
Pritam.khan@aiub.edu,diptagomes@aiub.edu

Abstract

An improved method to create an enhanced Bangla standard and local speech. The wav2vec 2.0 model has been fine-tuned using additional datasets collected alongside OpenSLR data. Our findings have shown that there are gains in transcription accuracy of as much as eleven percent, which is impressive given the low resources and languages employed, proving the merits of transfer learning and fine-tuning. The work of the research is aimed at expanding the knowledge base concerning the use of novel deep learning algorithms in small languages in the field of speech technology. The evaluation metrics included Word Error Rate (WER) and Character Error Rate (CER), with the fine-tuned model achieving an overall WER of 11.27% and CER of 6.03%. Comparative analysis with previous work shows a significant improvement from baseline models, highlighting the efficacy of the wav2vec 2.0 model in leveraging large and diverse datasets. The experimental setup was supported by a cluster computing environment with NVIDIA CUDA-compatible GPUs, underscoring the computational resources required for effective Automatic Speech Recognition (ASR) model training. The results demonstrate substantial advancements in ASR performance for Bengali, with the fine-tuned model outperforming previous benchmarks and showcasing the benefits of self-supervised learning approaches.

Keywords

Bangla Speech Recognition, wav2vec 2.0, Transfer Learning, Speech Technology, Automatic Speech Recognition (ASR).

1. Introduction

With the emergence of Artificial Intelligence, various facets of human life have improved, thus creating a scope of further exploration and research. In the field of Speech Recognition, Speech-to-Text (SST) and voice recognition is a well-established field of research addressing some important problems in daily life. Therefore, Speech-to-text (STT) is an essential domain of research that has huge impact in real life, including an aid to help hearing-impaired people follow through conversations, a method of detecting child abuse (Vásquez-Correa and Álvarez Muniain 2023), a tool for transcribing speeches into writings, and voice-controlled interfaces. Although there has been active development work in STT systems for large languages like English and Mandarin (Zhang, Haddow and Sennrich 2022), (Li et al. 2023) a lot of work remains to be done in the case of many Low Resource Language (LRL) (Sinha et al. 2024), including Bangla (Akther and Debnath 2022). More than 290 million people