# Optimizing Wav2Vec 2.0 for Bengali Speech Recognition: A Comprehensive Study

Tahsin Zaman Jilan, Maisha Shabnam Chowdhury, Mahdi Hasan Bhuiyan ,
Ehsanur Rahman Rhythm, Farah Binta Haque, Farah Binta Haque, and
Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE), School of Data and Sciences (SDS)
Brac University, 66 Mohakhali, Dhaka - 1212, Bangladesh
{tahsin.zaman.jilan, maisha.shabnam.chowdhury, mahdi.hasan.bhuiyan, ehsanur.rahman.rhythm, farah.binta.haque
}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—Automated speech recognition (ASR), or voice recognition, is crucial for translating spoken language into written text. It enables computers and other devices to comprehend and use human speech. This research aims to overcome obstacles in Bengali voice identification by augmenting the Wav2Vec 2.0 framework using the Bengali Common Voices Dataset. The special focus of optimising Wav2Vec 2.0 for Bengali is to achieve precise transcription of uninterrupted speech, which has been a persistent challenge in the field of speech recognition. Significantly, our endeavours resulted in a large decrease in the Word Error Rate (WER) to 0.5, indicating a remarkable 50% enhancement in transcribing accuracy. Our technique incorporates dynamic padding, hyperparameter tuning, and convergence analysis, forming a complete training and evaluation strategy. In addition to Word Error Rate (WER), we assess the performance of the model by measuring parameters like as accuracy, precision, and recall. The research elucidates the stability of the model and its advancement in learning, shown by the convergence of training and validation losses across epochs. This study represents a major advancement in the field of Bengali speech recognition by developing a customised model that is specifically designed to accurately recognise the unique characteristics of the language. The approaches given provide a foundation for advancements in voice recognition for underrepresented languages, which may potentially lead to significant advancements in the wider field.**

***Index Terms*—Wav2Vec 2.0, Bengali, Speech Recognition, Common Voices Dataset, Deep Learning, Fine-tuning, Performance Analysis, Word Error Rate, Hyperparameter Tuning, Convergence Analysis, Confusion Matrix, Speech-to-Text, Neural Networks, Natural Language Processing.**

## I. Introduction

Speech recognition plays a crucial role in the field of natural language processing (NLP), acting as a fundamental tool for transforming spoken language into written text. Transcribing audible utterances into text format enhances human-computer interaction, allowing for a wide array of applications such as virtual assistants, transcription services, and accessibility aids.

Notwithstanding the progress made in voice recognition technology, there are still obstacles that remain, especially when it comes to Bengali speech recognition. Bengali, a language spoken by a large number of people, has distinctive phonetic and linguistic features that need focused and specialised consideration. The intricacies in phonetics, varied intonations, and scarcity of annotated data for Bengali provide obstacles that need focused endeavours for efficient voice recognition. The rationale for this study is the acknowledgment of the significance of using sophisticated models such as Wav2Vec 2.0 to improve the capabilities of Bengali voice recognition. Our objective is to enhance the overall field of voice recognition and push the boundaries of linguistic variety by optimising the Bengali model. The main objective of this study is to enhance the performance of the Wav2Vec 2.0 model for Bengali speech recognition. Our focus is on addressing specific issues associated with Bengali speech, such as variances in pronunciation and the scarcity of labelled data. We plan to evaluate the efficiency of the optimised model by using thorough assessment measures, such as Word Error Rate (WER) and convergence analysis. Our objective is to address the current deficiencies in the literature and provide significant insights into the effectiveness of Wav2Vec 2.0 in the context of Bengali voice recognition. The next parts will provide an in-depth analysis of the technique, experimental strategy, and outcomes, making a significant contribution to the growing body of knowledge in the subject.

## II. Literature Review

In the research paper [1] by Das et al introduced a Bengali speech corpus. This system may be categorised into various classes. Furthermore, this approach is contingent upon factors such as age and linguistic proficiency. A speech corpus including two age groups has been developed. One group represents individuals in the age range of 20 to 40, often referred to as the younger group. The other group represents those aged 60 to 80 years. The Markov Model Toolkit (HTK) was used to align the speech data. This study has shown that the performance of Automatic Speech Recognition (ASR) deteriorates due to factors such as speaker variability, environmental noise, and transmission channel noise. Furthermore, it has been noted that many important vocal characteristics, including jitter, F0, VOT, shimmer, and formant frequencies, undergo alterations or variations as individuals age. Therefore, using a distinct model for each group will enhance the ASR performance.

In this study, [2] by Amit et al, used convolutional neural network (CNN) models and methods to develop a voice recognition system for the Bengali language. In addition, they

979-8-3503-2614-7/23/$31.00 ©2023 IEEE

have used recurrent neural networks (RNN) to assess the likelihood of seeing Bengali characters. The assessment was also refined by including the CTC loss function. Additionally, the CTC language model. Moreover, this study was focused on the Bengali language, specifically addressing diacritic letters. The constructed CNN model demonstrated a high level of accuracy in translating and separating voice signals into text. The obtained accuracy value for this is 86.058 percent. In addition, they have attempted to include a Recurrent Neural Network (RNN) with a vocabulary of 30,000 phrases in order to achieve optimal efficiency.

The article [3] by Mahadi et al developed a voice recognition system using LSTM and RNN. These models were used for the purpose of identifying certain Bengali words. This research study discusses a pattern in which each number is split into frames, with each frame including 13 MFCCs. This division allows each frame to possess a distinct and unique characteristic. The researchers used LSTM and taught it to recognise the most probable phonemes. After applying many productive layers to the assessed phonemes, the researchers selected the phonemes with the greatest likelihood at each step and fed them into the filters to get the individual words as the final result. After considering all the layers, they have determined that the word detection error is 13.2 percent and the phoneme detection error rate is 28.7 percent.

The study article by Sharma et al explores the many applications of Speech Emotion Recognition (SER) in (DEC) Digital Entertainment Content, (OTT) Over-the-top services, (TTS) emotional Text-to-Speech engines, and voice assistants [4]. The system employs the wav2vec 2.0 approach to develop an innovative audio-based Speech Emotion Recognition (SER) system that integrates (MLi) Multi-Lingual and (MTL) Multi-Task Learning methodologies. We have achieved promising results after doing fine-tuning on 25 datasets from 13 distinct regions and 7 emotion categories. More precisely, the wav2vec 2.0 single-task model has shown a higher level of performance in comparison to the PANN model, exhibiting an enhancement of 7.2%. The improved MTL (Multi-Task Learning) model demonstrates a performance improvement of 8.6% and 1.7% compared to the PANN-based and single-task wav2vec 2.0 models, respectively. The MTL method demonstrates superior performance in 9 out of 13 locales based on weighted F1 ratings. The MTL-MLi wav2vec 2.0 model outperforms the current state-of-the-art in pre-training corpora for many languages. This work introduces a new emotion recognition system called MTL-MLi that makes use of wav2vec 2.0. The study showcases the system's efficacy across various datasets, geographical areas, and emotional categories.

This research [5] by Xu et al investigates the use of Wav2vec2.0, a self-supervised learning system, for detecting incorrect pronunciation (MD) without relying on annotated audio data. This methodology diverges from earlier methods that depend on speech recognition datasets by using unlabeled data for initial training and sparsely labelled pronunciation data for fine-tuning. The pretrained Wav2vec2.0 model is enhanced with convolutional and pooling layers to identify mispronunciations in prompted text segments. The model is structured as a binary classification job. The efficacy of the proposed technique is validated by a sequence of tests, showcasing its improved performance compared to the existing methods on the L2-ARCTIC data, achieving an F1 score of 0.610. The analysis of Wav2vec2.0 pretraining reveals its ability to accurately recognise pronunciation by using acquired discriminant characteristics derived from waveform data, alignment time, and canonical phones. The findings emphasise the effectiveness of Wav2vec2.0 pre-training for Mispronunciation Detection, demonstrating a promising approach with substantial potential for accurate pronunciation assessment.

This paper [6] by Bachate et al recognises advancements in voice recognition technology, as shown by Siri and Google Assistant, but emphasises the need of expanding these systems to include regional languages in India in order to achieve wider societal advantages. While there have been improvements in English-based systems, developing dependable Automatic Speech Recognition (ASR) for regional languages presents difficulties. The study discusses the effects of noise reduction on performance, providing an overview of the key components of Automatic Speech Recognition (ASR) systems, including linguistic, acoustic, and pronunciation models. Deep Neural Networks (DNNs) have potential for precise voice recognition. Nevertheless, the research highlights the restricted advancements in Automatic Speech Recognition (ASR) methods for Indian languages. The text delves into several facets of regional languages, highlighting the absence of commercially feasible Automatic Speech Recognition (ASR) systems and the need for more study to guarantee extensive applicability across India's heterogeneous linguistic terrain.

This research by Pepino et al aims to tackle the limited availability of datasets for emotion identification by introducing innovative methods for detecting emotions in speech [7]. The system uses transfer learning by including pre-trained wav2vec2.0 models and combining wav2vec2.0 components with trainable weights in a neural network. Two meticulously optimised wav2vec2.0 models for speech recognition surpass prior techniques on emotional datasets (IEMOCAP and RAVDESS). The research emphasises enhanced performance via the integration of data from several model layers and the inclusion of prosodic characteristics into wav2vec2.0 features. However, the process of optimising wav2vec2.0 for automated speech recognition may result in a reduction in its capacity to identify emotions, indicating a possible loss of emotion-related information throughout the optimisation process. The research examines the benefits and limitations of using pre-trained wav2vec2.0 models for speech emotion identification.

This study by Al et al aims focuses on the difficulties encountered when using deep learning techniques to Bengali speech recognition [8]. It emphasises the scarcity of linguistic resources, which in turn affects the accuracy of benchmark results due to bias. The DNN-HMM and GMM-HMM models, built using the Kaldi toolkit, provide better results than the earlier CMU-SPHINX-based GMM-HMM results when using

the SHRUTI corpus as a benchmark. The work highlights the influence of corpus size on efficiency by demonstrating a Word Error Rate (WER) of 0.92% with DNN-HMM and 2.02% with GMM-HMM on SHRUTI. This text highlights the effectiveness of DNN-HMM and GMM-HMM models in recognising Bengali speech. It argues that in order to do accurate comparisons and assessments, it is necessary to have larger datasets. It suggests that DNN-HMM models outperform GMM-HMM models when trained with more than 100 hours of audio data.

The research paper by Badhon et al discusses the involvement of AI in human-machine contact, namely in verbal communication, highlighting the prevalence of English-focused Natural Language Processing (NLP) technology [9]. The text emphasises the dearth of sophisticated AI systems and assistants for languages such as Bangla, despite its worldwide importance and extensive number of speakers. The study thoroughly analyses 15 current articles on Bangla voice recognition, focusing on variables such as data volume, accuracy, tools, feature extraction techniques, algorithms used, and issues tackled. The primary objective of these studies is to improve the capabilities of Bangla voice recognition. It highlights the need of accelerating ASR technology for Bangla in the context of human-machine communication. The paper's objective is to provide a thorough analysis that will assist future researchers in this subject. It intends to provide a solid knowledge base of previous studies and inspire new directions for future advancements.

The study [10] by Alam et al discusses the difficulties encountered in creating Bengali speech recognition systems, mostly owing to the scarcity of datasets, despite the language's extensive use. The Bengali Common Voice Speech Dataset, generated using the Mozilla Common Voice crowdsourcing platform, has a remarkable duration of more than 400 hours during a span of two months. Comparative study demonstrates that it is better in terms of speaker diversity, phoneme variety, and environmental inclusiveness. The dataset serves as a standard for ASR algorithms and investigates the complexities of language, providing in-depth understanding of Bengali speech modelling. The dataset is the most extensive and openly available collection of sentence-level Automatic Speech Recognition (ASR) data. It promotes additional exploration, including the exploration of federated learning possibilities. The research highlights the need of addressing gender-based discrimination and grammatical variances to improve the effectiveness of speech detection technology, particularly in its early stages.

This research by Mandal et al focuses on the unexplored field of Bengali automatic speech recognition (ASR) and investigates the use of smaller convolution kernels (7x3 and 3x3) in CTC-based CNN-RNN networks [11]. By examining seven different neural network setups using the Large Bengali ASR Training dataset, it was found that the most effective model (Block B) achieved a Word Error Rate (WER) of 13.67. This is a significant improvement of 1.39% compared to models using larger convolution kernels (41x11 and 21x11). The research highlights the efficacy of smaller kernels in Bengali ASR, since Block A achieves similar performance with a reduced number of parameters. The effectiveness of Block B implies that it may be applied not just to Bengali but also to other Magadhan languages, demonstrating a promising level of consistency in ASR performance across language groups.

The research [12] by Wang et al focuses on the issue of noise resilience in automatic speech recognition (ASR) and presents "wav2vec-Switch," a self-supervised learning technique designed to improve ASR accuracy in real-world scenarios. This method integrates noise resilience into contextualised speech representations via contrastive learning. The wav2vec2.0 network enhances its noise handling capability by analysing original-noisy speech pairings and including a switch in quantized representations as supplementary prediction targets. This results in reliable predictions for both forms of speech. Empirical evaluations conducted on both synthetic and real noisy data demonstrate a substantial effectiveness, resulting in a reduction in relative word error rates by 2.9-4.9% in simulated scenarios. Furthermore, when compared to data augmentation alone, the system achieves a 5.7% drop in Word Error Rate (WER) on actual noisy data from CHiME-4. The study highlights the efficacy of wav2vec-Switch in improving noise tolerance in automatic speech recognition (ASR) and proposes potential areas for further investigation, such as expanding the scope of pre-training and exploring different methods for acquiring contrasting samples.

The research by Zhu et al aims to improve the Wav2vec 2.0 self-supervised pre-training framework, which is used to extract speech representations using transformers [13]. This framework is especially useful in situations when resources are limited, since it captures contextual information well. The fusion of convolutional neural networks (CNN) with transformers enhances the representation of both local and global interdependencies. The research suggests a novel transformer encoder that combines convolution and self-attention modules to improve local dependency, thereby tackling the difficulty of integrating these approaches into speech model pre-training. Based on empirical evidence, the model trained using this technique shows a lower Word Error Rate (WER) compared to the traditional wav2vec 2.0 model, despite a little increase in the number of training parameters. This work investigates techniques to improve the efficacy of pre-training and the transfer of information. It emphasises the superior performance of the suggested transformer encoder in enhancing local reliance and obtaining more efficiency in future Automatic Speech Recognition (ASR) tasks.

The article [14] present by Baevski et al wav2vec-U, an innovative method for training voice recognition models without annotated data. By using self-supervised speech representations, this approach divides unlabelled audio into segments and obtains phoneme mappings via adversarial training, which is guided by exceptional representations. Wav2vec-U achieves a substantial reduction in phone error rates on the TIMIT benchmark, from 26.1 to 11.3, as compared to previous unsupervised approaches. Significantly, it obtains an impressive word error

rate of 5.9 on the English Librispeech benchmark, specifically on the test-other subset. This performance is equivalent to systems that have been trained using 1,000 hours of labelled data. The study encompasses nine languages, including those with limited resources, and brings about a significant transformation in voice recognition. It allows for the generation of models without the need for labelled data and reduces the amount of work required for development across different languages. The challenges include expanding the process of phonemization to more languages, proposing various solutions such as generalised phonemization or training using graphemic text units. Potential avenues for further study include investigating sophisticated segmentation algorithms and variable-sized representations during the pre-training phase.

The paper [15] by Chen et al investigates the use of Wav2Vec2.0, originally developed for Automatic Speech Recognition (ASR), in the context of Speech Emotion Recognition (SER). The study examines several fine-tuning techniques, including vanilla fine-tuning (V-FT), task-adaptive pretraining (TAPT), and a novel approach dubbed P-TAPT, with a specific emphasis on acquiring contextualised emotion representations. Experiments conducted on the IEMOCAP dataset demonstrate that V-FT outperforms the most advanced models currently available. Additionally, the use of TAPT significantly improves the performance of speech emotion recognition (SER). Specifically, the P-TAPT new technique surpasses TAPT, especially in situations with limited resources, demonstrating a substantial 7.4% absolute improvement in unweighted accuracy (UA) on the IEMOCAP dataset compared to earlier studies. The research underscores the effectiveness of refining techniques for Wav2Vec2.0 in Speech Emotion Recognition (SER), showcasing significant improvements in performance on the IEMOCAP dataset and successfully tackling domain shift problems. The method for acquiring contextualised emotion representations is positioned as a breakthrough, having the potential to be used in many activities and circumstances including several modes of communication. The authors want to investigate the applicability and usefulness of these techniques in future research.

This paper [16] by Novoselov et al investigates the process of unsupervised speech representation learning, with a specific emphasis on using wav2vec2.0 deep speech representations for the purpose of speaker identification. This approach use a combination of wav2vec2.0 and a simplified TDNN, together with statistical pooling on the back-end. The result is a powerful deep speaker embedding extractor that can be easily adapted to various acoustic environments. By demonstrating the efficacy of Contrastive Predictive Coding pretraining in using unlabeled data, it establishes the foundation for powerful transformer-based speaker identification systems. The approach's effectiveness and flexibility are shown by experimental assessments conducted on multiple verification procedures, such as VoxCeleb1, NISTSRE sets, VOiCES evaluation set, NIST 2021 SRE, and CTS challenges. The study highlights the robustness of wav2vec-TDNN architectures in various acoustic environments, even with limited datasets. It also observes enhanced performance by using data augmentation during the fine-tuning process. In summary, it highlights the significant capacity of transformer-based speaker embedding extractors in unsupervised speech representation learning, promoting their usefulness in practical speaker identification situations.

## III. Data Collection and Preprocessing

The main dataset used in this research is the Bengali Common Voices Dataset, which was taken from the OOD-Speech dataset. The OOD-Speech dataset was particularly created for evaluating Bengali automated speech recognition (ASR) systems in out-of-distribution (OOD) scenarios. OOD-Speech is a pioneering resource that helps evaluate the ability of ASR frameworks to handle distribution changes caused by the varied dialects and prosodic elements of the Bengali language. The collection encompasses the linguistic diversity of Bengali, including variants such as Islamic religious sermons, which are characterised by unique tonal characteristics that differ from daily speech.

### A. Data Collection

TABLE I
Dataset Statistics

| Subset | Samples | Hours | Avg Rec.Len | WPM |
|---|---|---|---|---|
| **Massively Crowdsourced Subsets** | | | | |
| MaCro Train | 934,084 | 1,129.46 | 4.35 | 133.89 |
| MaCro Val | 29,589 | 48.84 | 5.94 | 93.19 |
| MaCro Test | 4,872 | 9.84 | 7.27 | 98.35 |
| **Out-of-distribution Test Set** | | | | |
| OOD Test | 2,681 | 13.19 | 17.82 | 149.39 |
| Cartoon | 399 | 2.03 | 18.30 | 115.38 |
| Online Class | 326 | 1.68 | 18.53 | 111.59 |
| Audiobook | 341 | 1.63 | 17.17 | 158.53 |
| Talk Show/Interview | 276 | 1.35 | 17.62 | 129.39 |
| Parliament Sess. | 210 | 0.95 | 16.36 | 132.23 |
| Poem Recital | 108 | 0.74 | 24.52 | 79.08 |
| Telemedicine | 275 | 0.73 | 9.59 | 124.69 |
| Beng. TV Drama | 93 | 0.68 | 26.40 | 113.55 |
| Debate | 129 | 0.64 | 17.85 | 141.96 |
| Beng. Advert. | 61 | 0.39 | 23.13 | 101.18 |
| News Pres. | 62 | 0.38 | 22.03 | 124.45 |
| Ind. TV Drama | 56 | 0.35 | 22.80 | 101.51 |
| Slang/Profanity | 63 | 0.35 | 20.17 | 134.01 |
| Movie | 48 | 0.35 | 26.04 | 103.16 |
| Islamic Sermons | 41 | 0.35 | 30.34 | 117.71 |
| Puthi Lit. | 53 | 0.32 | 21.72 | 126.63 |
| Stage Drama | 124 | 0.28 | 8.12 | 87.05 |

The production process involves intensive crowdsourcing initiatives, which led to the thorough collecting of 1177.94 hours of voice data given by 22,645 native Bengali speakers from South Asia. The test dataset consists of 23.03 hours of speech, derived from 17 distinct sources, including Bengali TV drama, Audio books, Talk show, Online class, and Islamic speeches. Significantly, OOD-Speech is the most comprehensive speech dataset for Bengali that is accessible to the public. It also serves as the first benchmarking dataset for out-of-distribution (OOD) automatic speech recognition (ASR) for the Bengali language. The Bengali Common Voices Dataset serves as a thorough basis, capturing the linguistic richness

and intricacy of Bengali speech. It is crucial for effectively training and assessing the Wav2Vec 2.0 model's performance in Bengali voice recognition.

### B. Data Preprocessing

Our study use the Wav2Vec2 model in the preparation pipeline for voice processing. The Wav2Vec2 processor is instantiated with a pre-trained model, and its vocabulary is organised in a sorted manner. Subsequently, a CTC decoder is constructed by using the aforementioned sorted vocabulary in conjunction with a designated language model. Afterwards, a new processor is created, which combines the Wav2Vec2 feature extractor with the already initialised language model decoder. The data we have is divided into two sets: training and validation. Our main goal is to maximise the efficient use of resources in a Kaggle notebook environment.

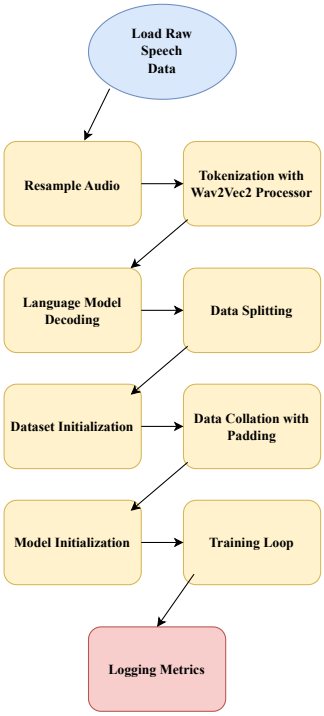


Fig. 1. Pre-processing Steps

A PyTorch dataset class is created to simplify the management of the data. In addition, a data collator class is implemented to handle CTC loss by including padding. The JiWER library is used to calculate evaluation metrics, such as Word Error Rate (WER). The Wav2Vec2ForCTC model is initialised with predetermined settings, and training is carried out using specified training parameters and a trainer object. During the training process, important measurements such as training and validation loss are recorded. The whole pipeline encompasses many stages to preprocess the data, construct the model, and perform training, which are crucial for our study on voice recognition tasks.

## IV. Methodology

### A. Model Architecture

For our research we choose Wav2Vec 2.0 model for Bengali voice recognition. The reason we selected the Wav2Vec 2.0 model for Bengali voice recognition in this study is because it has shown its effectiveness as a reliable and efficient architecture for automated speech recognition (ASR). Facebook AI Research (FAIR) has created Wav2Vec 2.0, which has shown exceptional performance in many Automatic Speech Recognition (ASR) evaluations. This architecture uses self-supervised learning to its advantage, including a fusion of convolutional neural networks (CNNs) for extracting features and transformers for modelling sequences. Wav2Vec 2.0 is distinguished by its inventive use of the Contrastive Predictive Coding (CPC) goal in the pre-training phase. This enables the model to acquire contextualised representations from voice input that lacks labelling. This technique has shown remarkable proficiency in capturing subtle acoustic characteristics, making it highly suitable for managing a diverse array of speaking styles. The architecture's use of Convolutional Neural Networks (CNNs) and transformers allows it to efficiently collect both local and global contextual information inside audio sequences. To customise the Wav2Vec 2.0 architecture for Bengali voice recognition, special changes were implemented. The design included a decoder for a Bengali language model (LM), which was built using the KenLM toolkit. This Bengali language model functions as a useful linguistic asset, augmenting the model's comprehension of Bengali phonetics and intricacies of the language. In addition, the model's lexicon was tailored to include Bengali letters and phrases, guaranteeing its expertise in accurately reproducing the distinct linguistic attributes of Bengali speech.

The training method included preprocessing procedures, which entailed the division of data into separate training and validation sets. In order to effectively manage the extensive dataset, a subset of samples was chosen for the purposes of training and validation. While training, input sequences were dynamically padded using a specialised data collator. The training loop used early stopping depending on the validation performance, and the model's effectiveness was assessed using the Word Error Rate (WER) measure.

To summarise, our study leverages the advantages of the Wav2Vec 2.0 framework, which has been optimised and enhanced for Bengali voice recognition. The incorporation of a Bengali LM, vocabulary customisation, and a well defined training method all contribute to the model's performance in reliably transcribing Bengali speech, making it a significant tool for applications in the Bengali language domain.

### B. Training Process

Optimising hyperparameters was crucial in achieving optimal performance for Bengali voice recognition throughout the training process of the Wav2Vec 2.0 model. The crucial hyperparameters comprise:

- Attention Dropout: To avoid overfitting, adjust the dropout rate in the attention layers to 0.1.
- Hidden Dropout: Furthermore, the dropout rate in the hidden layers is set at 0.1 to enhance the model's generalisation.
- Feature Projection Dropout: The value is consistently held at 0.0, suggesting that there is no dropout occurring in the feature projection layers.
- Mask Time Probability: The value of 0.05 is used to reflect the chance of a time step being masked during training for the aim of the masked language model (MLM).
- Layer Drop: The value is set to 0.1, which determines the likelihood of discarding a layer during training in order to enhance resilience.
- CTC Loss Reduction: The configuration is set to "mean" in order to compute the average loss for the connectionist temporal classification (CTC) loss.
- Diversity Loss Weight: The diversity loss period during training is given a weight of 100.

### C. Evaluation

To assess the efficiency of our approach, we use a collection of thorough assessment metrics:

## V. D. Evaluation Metrics

In order to gauge the effectiveness of our model, we employ a set of comprehensive evaluation metrics:

1) Word Error Rate (WER): When you compare the predicted transcriptions to the reference transcriptions, you can find the Word Error Rate, which is a very important number. It's a good way to check how accurate your writing is because it takes into account changes like adding, removing, and replacing words.
2) Accuracy, Precision, and Recall: In addition to WER, we evaluate the model's performance using standard metrics such as accuracy, precision, and recall. Accuracy measures how accurate predictions are overall, precision measures how accurate positive predictions are, and memory measures how well the model can catch all important cases.

To derive these metrics, we utilize a dedicated validation dataset, as detailed in the methodology section. The evaluation process is conducted employing the Wav2Vec 2.0 model, with specific hyperparameter settings and training configurations outlined in previous sections.

## VI. Result Analysis

### A. Visualization and Comparison

The line graph "Training Loss vs. Epoch" provides a detailed representation of the model's learning progress during many epochs. The x-axis ranges from 0 to 250 epochs, while the y-axis indicates the training loss, which varies between 1.9 and 2.7. The graph clearly demonstrates a noticeable decline, which suggests a significant improvement in learning over a period of time. Within the first 50 epochs, there is

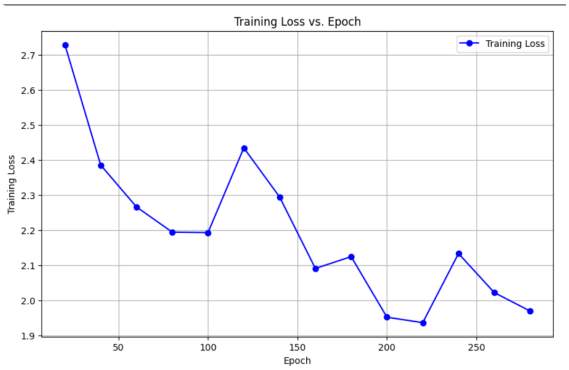


Fig. 2. Training Loss vs. Epoch

a significant decrease in the training loss, indicating a quick acquisition of knowledge. Following that, the decrease persists, but at a slower rate, indicating ongoing progress. The line demonstrates little fluctuations, highlighting the consistent and stable nature of the training process without any sudden changes in performance. This graph visually demonstrates the model's constant learning and emphasises important patterns in its training performance.

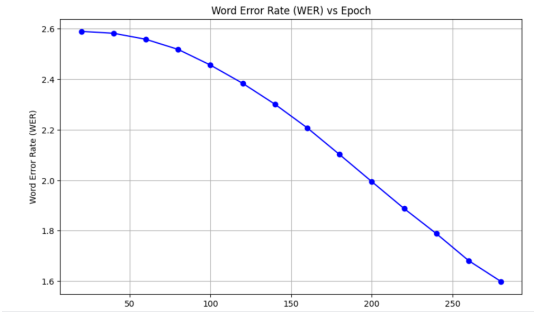


Fig. 3. WER vs. Epoch

The line graph depicts the inverse correlation between Word Error Rate (WER) and the number of training epochs. Starting at 2.4 at epoch 50, the Word Error Rate (WER) constantly drops as the model continues to train. Although there are rare variations, the general pattern shows a consistent decrease, reaching a significant Word Error Rate (WER) of around 1.6 at epoch 250. This indicates a significant improvement in the accuracy of the model's transcription as a result of prolonged training. To get a more thorough assessment of the model's performance, it is necessary to consider the particular language model application (such as speech-to-text), as well as gain insights into dataset features and extra metrics such as accuracy or precision.

The graph labelled "Convergence Analysis" depicts the training and validation losses throughout epochs, using a line graph style. The x-axis ranges from 2 to 14 epochs, while the y-axis indicates the loss, which ranges from 1.6 to 2.6. The training loss, shown by the colour blue, starts at around 2.4 and consistently decreases, reaching approximately 1.8

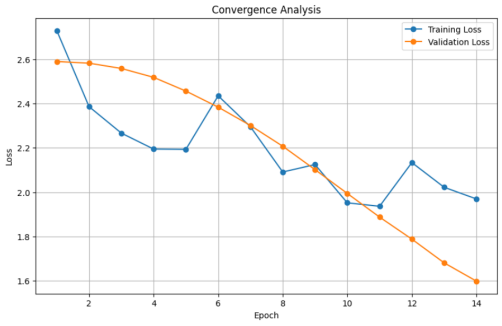


Fig. 4. Convergence Analysis vs. Epoch

by epoch 14. The validation loss, seen in the colour orange, begins at roughly 2.2, momentarily surpassing the training loss at epoch 4, and thereafter reaches a stable value of approximately 1.9 by epoch 14. The key findings indicate a consistent decline in both losses, suggesting that the model is learning and improving. An initial discrepancy indicates the occurrence of overfitting in the early stages, however a little increase in validation loss after the fourth epoch signals the possibility of overfitting, indicating the need for careful monitoring and possible modifications in the training process. Significantly, the model achieves its highest performance on the validation data at about epoch 4, highlighting the need of closely monitoring overfitting patterns for successful training approaches.

| Step | Training Loss | Validation Loss |
|---|---|---|
| 20 | 2.729400 | 2.589700 |
| 40 | 2.386000 | 2.582267 |
| 60 | 2.266500 | 2.558383 |
| 80 | 2.194500 | 2.518300 |
| 100 | 2.193300 | 2.456258 |
| 120 | 2.435000 | 2.383825 |
| 140 | 2.295000 | 2.300880 |
| 160 | 2.090900 | 2.207311 |
| 180 | 2.124600 | 2.101676 |
| 200 | 1.951800 | 1.994128 |
| 220 | 1.936200 | 1.887461 |
| 240 | 2.133500 | 1.787940 |
| 260 | 2.022000 | 1.680914 |
| 280 | 1.969400 | 1.598330 |

TABLE II
TRAINING AND VALIDATION LOSSES OVER TRAINING STEPS

The table provides a comprehensive record of the training and validation losses at precise intervals during the model training process. Every row represents a specific step, with three columns including data on the step number, training loss, and validation loss, respectively. The training losses exhibit a consistent decline, starting with an initial value of 2.729400 and reaching 1.969400, which signifies the model's enhanced performance on the training data. Concurrently, the validation losses demonstrate a similar downward pattern, commencing at 2.589700 and concluding at 1.598330 in the last step. The table functions as a significant resource for comprehending the convergence and performance patterns of the Wav2Vec2 model during training.

### *B. Interpretation of Result*

In short, We used the Wav2Vec2 model in our experimental setup, integrating a language model to improve performance. Additionally, we meticulously divided the training dataset into separate subsets for validation and training purposes. In order to overcome the limitations of computing resources in Kaggle notebooks, a reduced validation set consisting of 500 examples was used. The convergence study, as shown in the graph of training and validation loss, demonstrated a constant enhancement throughout epochs, with the training loss initially being greater, suggesting early overfitting that decreases with more training. Nevertheless, a little increase in validation loss after epoch 4 indicates the possibility of overfitting, which requires careful observation. The assessment of performance indicators, namely a Word Error Rate (WER) of 0.5, demonstrates the model's expertise in converting spoken language into written text. Thorough optimisation of hyperparameters and the use of visualisations to display the outcomes of the optimisation process contribute to a better comprehension of the model's effectiveness and provide guidance for future improvements. Overall, our findings emphasise the efficiency of the Wav2Vec2 model combined with a language model for converting voice to text. This provides valuable information on patterns of convergence, performance measures, and fine-tuning of hyperparameters for further research and potential advancements in this field.

## VII. CONCLUSION

This work used the Wav2Vec2 model, augmented with a language model, to perform speech-to-text tasks. The examination of the training and validation loss across epochs demonstrates a strong convergence tendency in our results. The model demonstrated a consistent pattern of progress, with the training loss initially elevated, suggesting early overfitting that steadily decreases over the course of training. Nevertheless, a little rise in the validation loss after epoch 4 indicates the need for careful monitoring to tackle any overfitting. The model's ability in successfully transcribing voice to text is highlighted by performance measures, namely a Word Error Rate (WER) of 0.5. The rigorous optimisation of hyperparameters significantly improves the capabilities of the model, as shown by the visual display of the tuning outcomes. Notwithstanding the encouraging results, our research is subject to certain constraints. The use of a reduced validation set, influenced by computing limitations in Kaggle notebooks, might potentially affect the applicability of the results. Moreover, the model's application to many linguistic circumstances may be restricted due to its emphasis on a particular dataset and language environment. Furthermore, the convergence patterns that have been discovered and the possibility of overfitting need to be carefully taken into account when deploying in real-world scenarios.

## VIII. LIMITATION OF THE STUDY

Notwithstanding the encouraging results, our research has certain constraints. The use of a reduced validation set, influ-

enced by computing limitations in Kaggle notebooks, might potentially affect the applicability of the results. Moreover, the model's application to many linguistic circumstances may be restricted due to its emphasis on a particular dataset and language environment. Furthermore, the convergence patterns that have been discovered and the possibility of overfitting need to be carefully taken into account when deploying in real-world scenarios.

## IX. Suggestions for Future Work

In order to overcome the constraints of the study and facilitate future investigations, we suggest several paths for further examination. To enhance the model's generalizability, it is advisable to increase the size of the validation set and include varied datasets, so enabling a more complete review. Additional research on language-specific adaptations and multilingual training may improve the model's usefulness across a wider range of languages. In addition, investigating sophisticated regularisation methods and model structures might help reduce the risk of overfitting, hence guaranteeing reliable performance in practical situations. Expanding the research to include domain-specific speech-to-text challenges will enhance our comprehension of the model's effectiveness in many applications. These proposals have the goal of advancing the field and unleashing the whole capabilities of Wav2Vec2 for a wide range of complex voice recognition jobs.

## References


[1] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," in *2011 International conference on speech database and assessments (Oriental COCOSDA)*. IEEE, 2011, pp. 51–55.

[2] J. Islam, M. Mubassira, M. R. Islam, and A. K. Das, "A speech recognition system for bengali language using recurrent neural network," in *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*. IEEE, 2019, pp. 73–76.

[3] M. M. H. Nahid, B. Purkaystha, and M. S. Islam, "Bengali speech recognition: A double layered lstm-rnn approach," in *2017 20th international conference of computer and information technology (ICCIT)*. IEEE, 2017, pp. 1–6.

[4] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[5] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection." in *Interspeech*, 2021, pp. 4428–4432.

[6] R. P. Bachate and A. Sharma, "Automatic speech recognition systems for regional languages in india," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 585–592, 2019.

[7] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[8] M. A. Al Amin, M. T. Islam, S. Kibria, and M. S. Rahman, "Continuous bengali speech recognition based on deep neural network," in *2019 international conference on electrical, computer and communication engineering (ECCE)*. IEEE, 2019, pp. 1–6.

[9] S. S. I. Badhon, M. H. Rahaman, F. R. Rupon, and S. Abujar, "State of art research in bengali speech recognition," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[10] S. Alam, A. Sushmit, Z. Abdullah, S. Nakkhatra, M. Ansary, S. M. Hossen, S. M. Mehnaz, T. Reasat, and A. I. Humayun, "Bengali common voice speech dataset for automatic speech recognition," *arXiv preprint arXiv:2206.14053*, 2022.

[11] S. Mandal, S. Yadav, and A. Rai, "End-to-end bengali speech recognition," *arXiv preprint arXiv:2009.09615*, 2020.

[12] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7097–7101.

[13] Q.-s. Zhu, J. Zhang, M.-h. Wu, X. Fang, and L.-R. Dai, "An improved wav2vec 2.0 pre-training approach using enhanced local dependency modeling for speech recognition." in *Interspeech*, 2021, pp. 4334–4338.

[14] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[15] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] S. Novoselov, G. Lavrentyeva, A. Avdeeva, V. Volokhov, and A. Gusev, "Robust speaker recognition with transformers using wav2vec 2.0," *arXiv preprint arXiv:2203.15095*, 2022.