Regional-Speech: A Bengali Speech Recognition Dataset for Benchmarking Models Under Dialect Variation in Puran Dhaka.

Submitted By
Naimur Rahman
211-35-3161
Department of Software Engineering
Daffodil International University

Supervised By
Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Daffodil International University

A thesis that was handed in to the Department of Software Engineering to complete the requirements for a B.Sc. in Software Engineering degree
Fall 24

APPROVAL

This thesis titled on "**Regional-Speech: A Bengali Speech Recognition Dataset for Benchmarking Models Under Dialect Variation in Puran Dhaka.",** submitted by **Naimur Rahman (ID: 211-35-3161)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

---------- **Chairman**

Professor Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **Internal Examiner 1**

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **Internal Examiner 2**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **External Examiner**

Dr. Md. Monowarul Islam
Associate Professor
Department of Computer Science & Engineering
Jagannath University

Supervisor Declaration

I hereby declare that I have checked this thesis in my opinion, this thesis id adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name: Md. Shohel Arman
Position: Assistant Professor
Date: 14 January 2025

Student's Declaration

It is hereby declared that

1. The thesis that was turned in was written solely by myself or within the course of completing the degree program at Daffodil International University.
2. So long as this has been informed appropriately through referencing, the thesis does not include any material that has been written and published before by someone different.
3. In the thesis no part has been submitted or accepted towards any other degree or diploma in any university or any other organization.
4. All major sources of assistance have been identified.

Naimur Rahman

(Student's Signature)

Full Name: Naimur Rahman
ID: 211-35-3161
Date: 14 January 2025

Ethics Statement (Optional)

We thus state that every assistant to this project respects the pro-active ACL Code of Ethics and is aware of it. Every operation carried out including human volunteers followed in consent.

Abstract

Regional dialects of the Bengali language, which is spoken throughout South Asia and among the Bengali diaspora, are influenced by historical, cultural, and geographic factors. Eastern Bengali, Manbhumi, Rangpuri, Varendri, and Rarhi are the five main dialects of Bengali based on phonology and pronunciation. There are additional differences in vocabulary, pronunciation, syntax, and morphology within Bangladesh. The distinctive characteristics of dialects from areas such as Dhaka, Chittagong, Sylhet, Rangpur, Rajshahi, Noakhali, and Barishal set them apart from both standard Bengali and from each other.

Notwithstanding this linguistic diversity, there is still a dearth of resources and research devoted to comprehending regional Bengali dialects and incorporating them into natural language processing (NLP) systems. By examining these dialects using thorough, data-driven linguistic analyses, such as phonetic and morphological studies, this study seeks to close this gap. We also evaluate the viability of creating computer models customized for these dialects, such as Automatic Speech Recognition (ASR) systems. Applications like virtual voice assistants and other Bengali language tools might be made possible by such models.

Our research aims to promote inclusivity and effective communication while advancing knowledge and supporting the preservation of regional Bengali dialects (Puran Dhaka “Dhakaiya”). In order to ensure the Bengali language's continued relevance in contemporary computational applications, this research helps to develop language technologies that respect the language's cultural and linguistic heritage by attending to the linguistic needs of Bengali-speaking communities.

Keywords: Automatic speech recognition; Regional Bengali speech; Wav2Vec2; Bengali Dialects; Linguistic Analysis; ASR; Dataset Benchmark; Dataset Curation

Acknowledgement

First and foremost, I want to express my gratitude to Allah, the Great Creator. for enabling me to travel this path and clearing the way of all significant roadblocks and obstacles while I finished my thesis.

Second, I would like to thank my supervisor, Assistant Professor Mr. Md. Shohel Arman, sir, who oversaw my thesis, for his guidance and assistance. Every time I contacted him with a question, he shared his knowledge and advice with me.

My parents, friends, and acquaintances come next. Without their prayers and thoughtful assistance, I could not have continued on this journey.

A big thanks goes Jafor Ahmed, Shaikat Barman, Asif and Roksana Akter for their help and support, Md. Rana, Saidul Islam, Najmul Nahid for helping with data and linguistic support. A special thanks to Tonmoy Shome for his excellent guidance and support.

And finally, a big thank to all the assigned data collectors and annotators shown in table 1 for their efforts.

Table 1: Name of all the data collectors and annotators

Hazaribagh	**Lalbagh**	**Suritola**	**Bongshal**
Data Collectors		**Data Collectors**	
Jafor Ahmed	Shaikat Barman	Shuvo Khan	Anik Neogi
Saidul Islam	Md. Rana	Shovon Rayhan	Najmul Nahid
Data annotators		**Data annotators**	
Naimur Rahman	Jafor Ahmed	Shaikat Barman	Saidul Islam

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human expression and society growth depend on language; hence, structured language is absolutely important. In computer science, developments in spoken and written language recognition have happened fast. This development seeks to improve accessibility in many spheres, including technological literacy and disability support [19].

Improvements in speech and writing recognition have inspired ideas in domains including language education [38], language disorder assessment [4], and agricultural support [48]. Notwithstanding these developments, the use of Speech-to-Text technology for regional variations of the Bengali language is still restricted mostly due to the lack of resources and available datasets.

Still, according to revised research, especially in relation to automatic speech recognition tasks, there is a notable discrepancy in both theoretical and computational linguistic studies concentrating on regional Bengali variations. Lack of available datasets and resources has not allowed current deep learning models to sufficiently address regional differences [24]. This research gap in this field led us to develop our research idea, which bridges these gaps by including an open-sourced dataset including many regional Bengali speech dialects and a modelling attempt acting as a fundamental resource for next research activities in this field.

Bengali is seventh most spoken language by the total number of speakers in the world, Bengali shows notable regional differences moulded by historical, geographical, and cultural influences. Reflecting a feeling of community identification and comfort among speakers, these dialectal variations cover vocabulary, pronunciation, grammar, and cultural quirks.

1. **Vocabulary:** Standard Bengali language employs a consistent set of words derived from both modern use and classical literature. On the other hand,

Regional dialects sometimes have local slang and expressions not found in formal or official settings.

2. **Pronunciation:** Bengali spoken in standard form and its regional dialects can have rather different tones. Along with unique phonetic features, variations in vowel and consonant pronunciation define every dialect.

3. **Grammar:** The basic grammar of Bengali remains fairly uniform across different dialects, but there can be variations in how particles are used, verb forms are conjugated, and sentences are structured. Some dialects might simplify or modify certain grammatical rules.

4. **Cultural Influences:** Regional dialects often reflect local cultures and tra- ditions, incorporating local sayings, idioms, and words borrowed from neigh- boring languages.

5. **Usage:** Standard Bengali is used in school, the news, and official documents, among other places. Regional dialects, on the other hand, are more common in everyday conversations within communities and social groups. This shows that people use different types of language depending on the situation.

6. **Writing and Literature:** Formal writing, including newspapers, books, and academic work, typically uses standard Bengali. Regional dialects are rarely used in formal writing but might appear in casual communication like social media.

7. **Geographic Variation:** Different areas in Bangladesh and India have their own dialects, shaped by local cultures, histories, and landscapes.

It is difficult to create and apply language policies supporting standard Bengali while yet honoring regional dialects. This process calls for a deliberate approach considering social, cultural, and linguistic aspects. Language changes organically; changes in the standard form may encounter opposition. Choosing which modifications to approve or reject calls for constant communication and consensus building.

Bengali's great linguistic variety should be acknowledged if one is to create an efficient implementation model. Many current datasets mainly focus on formal, standard

Bengali, which can limit users' ability to communicate in their natural style. This forces users to adapt to a more formal language, which can impede genuine communication. The distinct morphology and various accents within Bengali further complicate creating datasets that truly represent everyday language. Therefore, having large, diverse datasets is crucial for training comprehensive deep-learning models.

Lack of resources including automatic speech-to-text systems for Bengali dialects, text-to-speech systems including regional accents, and tools for dialect classification and transliteration projects has hampered research and development progress. Lack of these resources creates significant challenges to progress these technologies.

Accents are pronunciation differences between speakers of the same language, whereas dialects are linguistic characteristics unique to specific groups or regions. Geographical, cultural, and historical factors all contribute to these differences. To effectively recognise and process these accents, computational systems must first develop models that can accurately distinguish between them. This is critical for tasks such as identifying speakers, recognizing emotions, and assessing stress levels.

Alrehaili et al. [23] worked on analyzing Arabic dialects through a system that identifies dialects in audio recordings. They used a dataset of 672 audio samples from eight Arab dialects and applied Convolutional Neural Networks (CNN) for classification, achieving an accuracy of 83%. This work helps improve communication and translation across Arabic dialects.

Miao Wan et al. [50] focused on recognizing Chinese dialects using deep neural networks, with a special emphasis on regional accents. Their models achieved ac- curacies of 79.96% and 83.59%, enhancing applications like customer service and translation by addressing regional linguistic diversity.

Research on Bengali dialect recognition is still under progress in comparison. Among the few who tackle this area is Tomal et al. [49]. Their main concentration was on spotting regional Bengali dialects, mostly spoken but not well recorded. They attained a high accuracy of 96% by means of several data processing methods and analysis of large amounts of two dialects, Chatgaiya and Pabna. Their efforts draw attention to the need of appreciating and knowing the several dialects spoken in Bengalis. Though great efforts are being made to compile datasets, current resources suffer in speech diversity, size, and accessibility. To address issues involving regional

languages; hence, it is imperative to create specific materials catered to these requirements.

The SHRUTI read-speech corpus, released by IIT Kharagpur in 2011, comprises 21.64 hours of audio containing 7383 unique Bengali phrases and 49 phonemes. The IARPA voice corpus, associated with the Babel initiative, comprises 215 hours of Bengali telephone conversations and scripted communications from 2011 and 2012, along with accompanying transcripts [13]. In 2019, the Linguistic Data Consortium for Indian Languages (LDC-IL) acquired 138 hours of continuous Bengali speech and subsequently developed a series of speech corpora, as documented in a 2020 research paper [57]. The Technological Development for Indian Languages Initiative (TDIL) offers access to Bengali vocabulary through more than 43,000 audio recordings from 1,000 native speakers in West Bengal. The European Language Resources Association (ELRA) provides a 70-hour Bengali speech corpus [18], yet details regarding the publication date and corpus specifics are not disclosed or reviewed. In 2018, Khan et al. developed a connected-word speech corpus consisting of 62 hours of recordings from more than 100 speakers [40]. In that year, Khan and Sobhan created a second corpus centred on isolated words, comprising 375 hours of recordings from 150 speakers [3]. In 2018, Google released the "Large Bengali ASR training dataset" (LB-ASRTD) for the LVCSR challenge, providing a substantial Bengali speech corpus, which is available on the Open SLR website (Open SLR LB-ASRTD, 2018). The dataset comprises orthographic annotations in Bengali script and encompasses more than 229 hours of dialogue from 505 native Bangladeshis (323 men and 182 women), resulting in a total of 217,902 utterances.

In 2020, Ahmed and Sadeq [24] constructed a 960-hour annotated speech corpus using publicly available audio and text data and proposed a method for automatic transcription generation from existing audio recordings. SUBAK.KO [45] is a Bengali speech corpus originating from Bangladesh, consisting of 241 hours of recordings from local speakers distributed across eight divisions and thirty-four districts. Data were obtained from online platforms, including YouTube.

In 2022, Bhogale et al. [18] developed the Shrutilipi ASR Corpus, a labelled ASR corpus sourced from news broadcasts in 12 Indian languages, including Bengali. The dataset encompasses approximately 6400 hours of data, with the Bengali segment accounting for 443 hours. In 2023, Fazle et al. [24] and Bengali initiated a significant effort to support and advance research in this specific domain.Artificial intelligence resulting in

the Domain Diversified Bengali Speech Corpus, designated OOD-Speech, serves as a substantial dataset for Bengali speech recognition, specifically designed for out-of-distribution benchmarking. The dataset comprises Bengali sentences derived from over 2000 hours of transcribed audio recordings gathered from India and Bangladesh. The initiative seeks to compile 5000 hours of audio data by 2023, targeting critical linguistic challenges in Bengali speech recognition. This corpus is a key part of an initiative aimed at collecting 5000 hours of carefully gathered audio data by 2023. This study conducts a thorough analysis of the significant linguistic challenges involved in the development of Bengali speech recognition systems. Additionally, we present analyses and observations based on the extensive data available to us.

The study also presents a benchmark analysis of the Word Error Rate (WER) for an ASR algorithm based on Hidden Markov Models and Gaussian Mixture Models (HMM-GMM). The research is adaptive, subject to revisions, and integrates the most recent standards from the Common Voice Speech Corpus.

In summary, the key contributions of this research to the field include,

1. Leading the incorporation of an innovative aspect in the development of natural language processing for the Bengali language. This initiative seeks to improve the understanding and analysis of Bengali from computational and linguistic viewpoints, targeting deep learning engineers, researchers, and linguistic scholars.

2. Creation of a comprehensive and standardised research pipeline with protocols to direct the entire research process. This framework is designed for similar research initiatives and is expected to be applied in various Bengali-speaking regions in future studies.

3. Development of a meticulously crafted and curated 39-hour open-source speech corpus. This corpus is meticulously designed to capture subtle regional speech patterns, including diverse dialectical forms and variations, to facilitate research and development in speech recognition and dialectal studies..

4. Creation of a precisely calibrated model proficient in accurately transcribing regional Bengali speech. The model incorporates various dialects, speech pattern variations, and distinct vocabulary, while rigorously following established orthographic standards.

5. Performing an extensive linguistic analysis concentrating on regional variations of the Bengali language throughout various geographical areas in Bangladesh. This analysis seeks to elucidate linguistic diversity and evolution within the Bengali language context.

The remainder of this paper is structured as follows:
Chapter 2 offers an exhaustive analysis of the current research and development in this field. Chapter 3 elucidates the pertinent background study related to the subject matter. Chapter 4 delineates the methodology utilized in the research. Chapter 5 provides an analysis of the results, while Chapter 6 presents concluding remarks and summarises the study's findings.

Chapter 2

Literature review

The most basic form of human communication is speech, which also stands alone free from literacy restrictions. Developments in the field of speech recognition could help to democratize world knowledge and services availability. Through spoken language, these developments enable people of different literacy levels and language competencies to easily interact with digital platforms, access instructional resources, and use basic services by means of seamless interaction with technology. By closing digital gaps and encouraging more general participation in the digital age, research in speech recognition not only makes previously inaccessible technologies accessible but also improves task efficiency and user experiences, fostering society equity. [11].

This chapter reviews current speech recognition research with an eye towards available corpora and modelling strategies.

2.0.1 Speech Recognition Corpus

2.0.1.1 International speech corpora

Shivaprasad et al. [11] created a database for the Telugu language to enable a more thorough knowledge of its dialects, addressing the lack of stan-directed tools for dialect study in speech recognition. Using Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), they identified these dialects depending on speech patterns. Their results showed that in this situation the GMM method performed better than HMM. MengEn Zhai et al. concentrated on the Yulin dialect, a less-known regional dialect in China with insufficient data for speech recognition re-analysis [22].

They first gathered and built a single, this particular dialect focused speech corpus, then examined its phonological patterns and created a matching wordlist or dictionary. In low-resource dialect environments, this approach remarkably improved the speech recognition accuracy of the Yulin dialect by 15.42% over conventional approaches.

2.0.1.2 Speech recognition modeling approaches for different languages

S. Darjaa et al. [43] developed a computer program to distinguish between the accents of the Slovak language and its regional dialects. A substantial collection of recordings from speakers with various accents was amassed, resulting in a specialized database designed to enhance the program's accent recognition capabilities. Furthermore, the system was trained to identify standard Slovak spoken in a neutral tone. The findings demonstrated that the system effectively identified the primary accent groups in Slovak, highlighting its practical utility. The researchers indicate that additional data and advanced techniques may facilitate further improvements.

Imaizumi and Masumura [9] created a computer model for the recognition of various Japanese dialects. Owing to insufficient data, both standard and dialect-specific datasets were employed for model training. Initially, the two datasets were combined; however, this method led to the omission of dialect-specific details. They subsequently employed a supervised learning method, resulting in a 19.2% reduction in the error rate.

Swiss German poses considerable challenges owing to its varied writing styles and absence of standardization. Nigmatulina et al. [10] examined the understanding of Swiss German among different dialects. The researchers utilized two methodologies: dialectal writing, which involves the phonetic transcription of Swiss German, and normalized writing, which transcribes Swiss German to align with standard German conventions. The researchers evaluated these methods through a computer program utilizing a dataset that includes 14 distinct varieties of Swiss German. The normalized writing approach demonstrated superior performance regarding word error rates, whereas the dialectal writing approach was more effective in recognizing individual letters. This study enhances the algorithmic comprehension of Swiss German and suggests possible applications for other languages that lack standardized orthography.

Nguyen et al. [2] developed a tonal variation recognition system for standard Vietnamese, especially Hanoi. Wavelet transforms helped analyze large speech dataset pitch variations. Their Hidden Markov Models (HMMs) recognized tones better with one of the two monotonic references. They started with tonal recognition to understand Vietnamese words.

A Moroccan Darija speech recognition system was presented by Abderrahim Ezzine and his team at ISCV 2020 [7]. They tried to identify Moroccan Darija's first ten Arabic numbers. Their system used MFCCs and HMMs. They reached 96.27 recognition accuracy after many trials.

Low-resource language Automatic Speech Recognition (ASR) was proposed by Omar Aitoulghazi et al. [16]. They achieved 22.7% WER and 6.03% CER using Baidu's advanced "Deep Speech 2" model and 24 hours of spoken language data. They show that the system is not perfect, but they are promising for understanding the widely spoken Moroccan dialect.

The Arabic dialect ASR model was improved by Hamzah A. Alsayadi et al. [43]. Mixed CNN and LSTM networks with attention-based encoder-decoder techniques created a hybrid model. They created another language model and tested it on two Arabic dialects using RNNs and LSTMs. Their method had a 57.02 WER in experiments.

Jooyoung Lee et al. [38] say lack of research and resources makes Korean dialect recognition difficult. Their novel approach focused on intonation, not syllables. A hybrid model using fundamental frequency was trained using a BILSTM network with an attention mechanism. Even with non-dialect parts, this method found dialect-rich speech. Testing this model across various ages and speech styles, they achieved an accuracy of 68.51

The aim of the study carried out by Vivek Bhardwaj et al. [12] was improving ASR systems for Punjabi dialects. Pitch acoustic characteristics allowed their improved system to be tested using the Malwa and Majha Punjabi dialects. Though there is always room for development, the results showed that the system reflects improvement in knowledge of these dialects with WERs of 23.25% and 25.91%, respectively.

N. D. Londhe et al. assigned an ASR system to the rare Indian dialect "Chhattisgarhi" based on two widely used machine learning methods: artificial neural networks (ANN) and Support Vector Machines (SVM [40]). They assessed these methods using a dataset comprising fifty unique words uttered by fifteen individuals. We evaluated ANN and SVM against the conventional Hidden Markov Model (HMM). They also assessed the system's capacity to detect variances of the same words used by several speakers. Regarding this dialect, the ANN and SVM techniques seem to be reliable and successful.

2.0.1.3 Bengali speech corpora

Given its major impact on the academic and commercial sectors, the speech recognition field has attracted a lot of research attention in Bangladesh. Over time, a range of datasets have been made available or actively exploited for speech recognition model training. Among the several degrees at which these datasets are painstakingly annotated are phoneme, word, utterance, and sentence levels. They achieve both the twin objectives of simplifying model development and raising awareness of the linguistic variety and richness found in the languages spoken in these regions.

Although Bengali, sometimes known as Bangla, is still classified as a low-resource language since there are few publicly accessible resources for research and development dedicated to this linguistic area [13].

Developing very accurate and useful models for a variety of uses depends fundamentally on a large training dataset. Bengali speech recognition lacks many resources, thus there is a significant challenge as well. Although there are plenty of datasets, it is noteworthy that most of the larger and maybe more extensive ones are still kept in private domains and have not been made publicly available.

Surprisingly, there are shockingly few datasets spanning more than a thousand hours of recorded Bengali speech that sufficiently vary in terms of linguistic content and context. Emphasizing how pronunciation changes resulting from accents can affect ASR performance, Shafkat Kibria et al. [10] investigate in their research article the effect of regional accents on the accuracy of ASR systems. With special attention to vowels like monophthongal and diphthongal, the study concentrates on two different groups of speakers analyzing the acoustic features of their accents concerning Bangladeshi Bengali. The study covers speakers from the Sylhet area, known for their strong dialect differences, both male and female as well as people from other Bangladesh districts with rather milder dialect variations.

To classify and separate accents, the study looks at acoustic characteristics including pitch slope, formant frequencies, and vowel length. The results show appreciable variations in formant frequencies and pitch slope steepness among accents, which adversely affect ASR performance. The paper emphasizes the need of accent focusing acoustic models in order to more suit speakers from different dialect areas. It also emphasizes the need of including accent-related speaker variability in corpora development to enhance ASR systems for Bangladeshi Bengali.

For Bengali, there are few speech corpora accessible as well as open-resource. Among linguistic resources, the SUVAK.KO speech corpus [45] stands out as a clear exception. Designed for research in automatic speech recognition, this publicly available annotated Bangladeshi standard Bangla speech corpus Comprising 241 hours of outstanding quality speech data, this corpus consists of 229 hours for read speech and broadcast speech data respectively. Conducted in a controlled studio environment, the read speech recordings are in standard Bengali and feature contributions from 33 male and 28 female, representing 8 divisions and 34 districts of Bangladesh. The section on the read speech also consists of one hour and thirty minutes of recorded speech from two speakers learning two second languages, L2. A fraction of the dataset came from well-known internet sites including Face-book and YouTube. This method guarantees that the dataset conforms with modern communication techniques and reflects the use of modern language. Every part of SUVAK.KO has been painstakingly manually annotated to guarantee dependability and accuracy of the given labels.

Table 2.1: Available Speech Corpus for Bengali. Stats of the datasets taken from [40] mostly.

Year	Corpus Name	Size of Dataset	No. of Speakers	Publicly Available
2011	SHRUTI [6], [7], [9]	21.64 hours	26 males, 8 females	Yes.
2012	IARPA-babel103b-v0.4b [17]	215 hours	Not known	Not Publicly Available. Access per application.
2014	LDC-IL [27]	138 hours	240 males, 236 females	No.
2014	TDIL [22]	43,000 audio files	1,000 native speakers	Not Publicly Available. Available for TDIL members.
2018	OpenSLR [21]	229 hours	323 males, 182 females	Not Publicly Available. Accessible under Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0 US)
2018	Bengali Connected Word Speech Corpus [20]	62 hours	50 males, 50 females	Not known
2018	Bengali Isolated Word Speech Corpus [19]	375 hours	50 males, 50 females	Not known
2018	OpenSLR [21]	229 hours	323 males, 182 females	Publicly Available under Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0 US)
2019	ELRA [18]	70 hours	Not known	Not Publicly Available. Available for ELRA members.
2020	Bengali Speech Corpus from Publicly Available Audio & Text [24]	960 hours	268 males, 251 females	No
2020	Subak.ko [45]	241 hours	33 males, 28 females	Yes
2022	Shrutilipi [18]	443 hours	All India s Radio archives	Ye
2022	Common Voice Bengali Corpus [32]	1000 hours	22.1k Speakers	Not Publicly Available. Accessible under

Among the limited publicly accessible resources, the most prominent is the OpenSLR dataset developed by Google [48]. However, this dataset primarily focuses on the utterance level (Table 2.1) of speech recognition, leaving a considerable gap in broader linguistic and contextual coverage.

Even among speakers of the same language, accents represent differences in word pronunciation. While dialects cover variations in pronunciation, vocabulary, and grammar, accents mostly consist in differences in pronunciation. Dialects thus reflect a more general and more complete linguistic concept; accents are only one feature of total dialectal variation.

As a language, Bengali shows amazing variation with many separate regional dialects. Clearly, depending on the area—including several Bangladesh districts—these dialects have different ac-cent and pronunciation. Among areas including Dhaka, Chattogram, Sylhet, Noakhali, Barishal, and Rangpur one can find several dialects. Fascinatingly, depending on the particular area, even one word might change in pronunciation [93].

The creation of a good ASR system mostly depends on the acquisition of a large and varied training set that totally reflects the linguistic differences inherent in the target tongue. Strong enough data is required to let the ASR system correctly identify and transcribe speech, including the rich tapestry of regional dialects and accents inside the language territory.

With the resources available now for the Bengali language, regretfully there is a major obstacle on this road to ASR system excellence. Though valuable, Bengali speech data usually has a rather small scope and scale. This restriction makes it difficult for it to sufficiently show the whole spectrum of the several dialectal nuances and regional language variants defining the linguistic variation of the Bengali language.

This lack of readily available resources determines whether ASR systems built in the complexity of the Bengali language are reliable and accurate. Demanding ASR tasks including speaker-independent speech recognition or speech phrase detection in noisy acoustic environments highlight this difficulty especially.

Maximizing the present resources will help to eliminate these obstacles and enable the development of speech recognition technologies able to efficiently process spoken Bengali language. This augmentation should consist in generating a more complete, representative Bengali speech corpus. These improved language datasets would provide the basis for creating creative ASR technologies, ensuring that advanced speech recognition systems could correctly exploit and grasp Bengali linguistic richness and diversity.

2.0.2 State-Of-The-Art (SOTA) Speech Recognition Systems

The development of current deep learning technologies has sparked notable expansion in the field of speech recognition. The accuracy of the speech recognition system has improved rather significantly in recent years. By allowing the development of more complex models capable of learning from vast speech datasets, deep learning algorums—especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have transformed speech recognition tasks.

Unprecedented degrees of speech recognition accuracy have been reached by voice recognition systems made possible by deep learning algorithms. These modern models can now achieve precision approaching human-like accuracy by word recognition rates above 95%. This higher accuracy has greatly enhanced the dependability of speech recognition systems and created many new opportunities for the use of speech recognition technology.

One of the best illustrations of the progress in speech recognition is the virtual assistants included in common appliances. Task completion including note-taking, contacting people, or setting reminders involves Google Assistant on Android, Siri on iOS, Amazon's Alexa, and Windows' Cortana. Thanks to the advanced development of voice recognition technology, these virtual assistants can sense voice commands and react intuitively and spontaneously.

Recording and tak-king notes during meetings, translating and transcribing, supporting video game interactions, and automating closed captioning for video indexing are other places where speech recognition is quite important. Many businesses provide such services via apps or software.

Research in Automatic Speech Recognition is on creating systems using models to transcribe speech into text or other symbolic representations. Real-world uses for ASR models abound: they run voice-activated devices, translate languages, and create transcriptions for those with hearing loss.

When constructing an ASR model, it is essential to consider several critical factors:

1. **Data collection and preparation:** ASR models call for large audio datasets precisely reflecting the target language and its regional dialects. These sets have to be transcribed and labeled for use in evaluation and training. Common Voice [5], Voxceleb [23], and LibriSpeech [12] are among often used datasets.

2. **Feature Extraction:** Usually, ASR models run on spectral frequencies methodically obtained from the raw audio waveform. These characteristics help to accurately and efficiently recognize and interpret spoken English.

3. **Acoustic Modeling:** This core component of an ASR system transforms audio information into textual output. Intricate statistical models map the extracted acoustic features from the input audio data to the corresponding textual representations. Deep Neural Networks (DNNs), known for their ability to model complex relationships within data, are frequently used for acoustic modeling in ASR systems.

4. **Language Modeling:** Language models estimate in the target language the likelihood of various word combinations. By including past language knowledge into the recognition process, they increase the coherence and accuracy of ASR output. This entails approaches including advanced neural network-based language models using deep learning techniques to capture complex linguistic patterns and relationships and n-gram models, which analyze word sequences up to "n's" words in length.

5. **Evaluation:** ASR models are typically evaluated using metrics such as WER or CER, which measure the degree of deviation from the reference transcripts. These metrics express the error rate as a percentage of incorrect words or characters.

Some of the top-performing models in the field include Whisper [19], Jasper [22], Wav2vec 2.0 [26], Conformer CTC [8], Kaldi ASR [10], and Google Assistant [25].

2.0.2.1 Wav2Vec 2.0:

Wav2Vec 2.0, developed by Facebook AI [56], represents a sophisticated ASR system. Its primary objective is to transcribe speech into written text accurately. The architecture is illustrated in Fig 2.1.

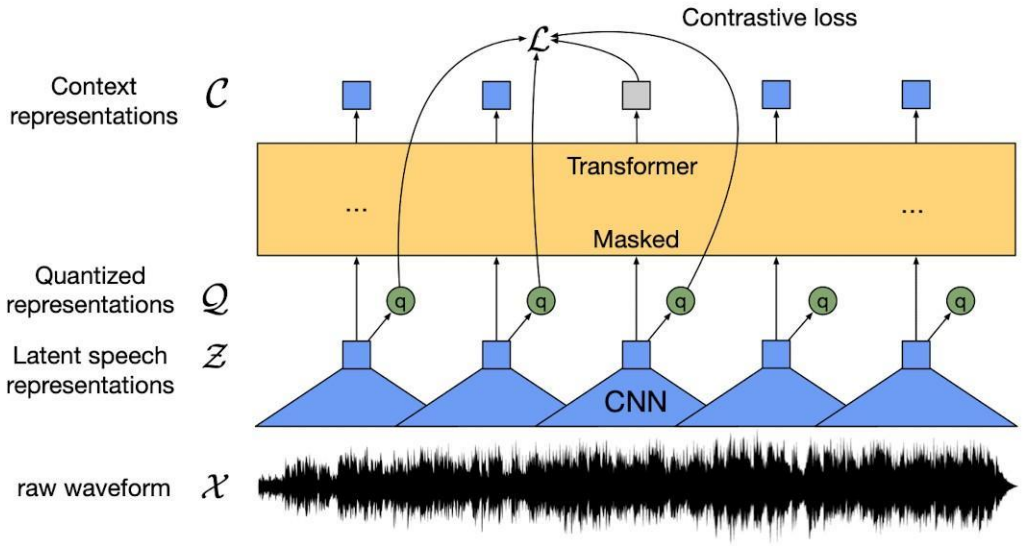


Figure 2.1: Illustration of the Wav2vec2 framework

Using a novel self-supervised learning method that lets the model be trained without any transcribed speech data, enabling the system to be effectively applied to speech data with limited labelling, Wav2Vec2.0 is particularly flexible.

Wav2Vec 2.0's fundamental method is based on a large corpus of unlabelled speech data pre-training a neural network. Later on, a smaller, labelled dataset allows this pre-trained network to be tuned especially for a given ASR assignment. On many speech recognition tests, this method has displayed rather good performance.

Wav2Vec2.0 represents a significant advancement in ASR technology since it reduces the need for manual transcription efforts and increases the adaptability over many languages and sectors. Its wide spectrum of uses covers voice assistants and transcription services, thus it is the ideal instrument for the evolution of speech recognition systems.

Wav2Vec2.0 distinguishes itself with its self-supervised learning method, which replaces transcribed audio data with elimination of need. For languages and dialects lacking clear defined data, this is especially useful.

Pre-training a neural network forms Wav2Vec 2.0's central idea

on a vast amount of unlabeled speech data. This pre-trained network is then fine- tuned with a smaller set of labeled speech data to perform specific ASR tasks. This methodology has demonstrated outstanding performance across various speech recognition benchmarks.

Wav2Vec 2.0 represents a significant advancement in ASR technology by reducing dependence on manual transcription efforts and enhancing adaptability across mul- tiple languages and domains. Its applications are broad, ranging from transcription services to voice assistants, making it a valuable tool in the development of speech recognition systems.

The feature encoder output is discretized into a set of finite speech representations using product quantization [29], facilitating self-supervised training.

The Gumbel-Softmax technique facilitates the selection of discrete codebook entries in a fully differentiable manner [13], [11], [16]. The hard Gumbel-Softmax operations *G* [16] and the straight-through estimator [1] are employed to achieve this. The output of the feature encoder is projected onto $l \in R^{G \times V}$ logits, and the equation below delineates the probabilities of selecting the *v*-th codebook entry for group *g*.

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(l_{g,k} + n_k)/\tau} \tag{2.1}$$

In the equation 2.1, τ is a non-negative temperature, $n = \log(-\log(u))$ and u are uniform samples from $U(0, 1)$.

During the neural network training of the pre-trained model, codeword i is chosen by $i = \arg\max_j p_{g,j}$, and the true gradient of the Gumbel softmax outputs are used respectively in the forward and backward pass.

The pre-trained model has learned to represent the speech audio by optimizing two types of losses during its training, which are,

1. **Contrastive loss (L_m):** Distinguishes the correct quantized latent speech representation from a set of distractors to learn representations.

$$L_m = -\log \frac{\exp(\mathrm{sim}(c_t, q_t)/\kappa)}{\sum_{q^{\sim} \in Q_t} \exp(\mathrm{sim}(c_t, q^{\sim})/\kappa)} \tag{2.2}$$

in 2.2, c_t is the context network output centered over masked time step t, q_t is the true quantized latent speech representation in a set of $K + 1$ quantized candidate representations $q^{\sim} \in Q_t$ which incorporates K distractors along with

q_t and sim(c_t, $q\tilde{}$) is the cosine similarity between context representations c_t and quantized latent speech representations $q\tilde{}$.

2. **Diversity loss (L_d):** Promotes diversity in the learned representations and to utilize all the codebook entries uniformly.

$$L_d = \frac{1}{G \cdot V} \sum_{g=1}^{G} (-H(\bar{p}_g)) = \frac{1}{G \cdot V} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v} \tag{2.3}$$

I equation 2.3, Each of the G codebooks has V entries. This is done by ensuring the utilization of the maximized entropy of the averaged softmax distribution l for each codebook entry $\bar{p}_g$ in an utterance batch. Note that the grumble noise is absent from the softmax distribution.

So, the total loss L becomes,

$$\mathrm{L} = L_m + \alpha L_d \tag{2.4}$$

In equation 2.4, α is a tuned hyperparameter to optimize the significance of the two losses. Upon showcasing the extraordinary performance of Wav2Vec 2.0 on the widely recognized English ASR dataset such as LibriSpeech [12], Facebook AI launched a multilingual variant known as XLSR (cross-lingual speech representations). This model extends Wav2Vec 2.0's capabilities by enabling the acquisition of speech representations that are beneficial across multiple languages.

In November 2021, Arun Babu et al. released XLSR's successor, XLS-R (short for 'XLM-R for Speech') [12] which was pre-trained using audio data spanning 128 languages that was almost half a million hours in duration and the model is available in model sizes ranging from 300 million to 2 billion parameters.

Fine-tuning is done by adding a single layer on top of an existing pre-trained network and then train the model using own custom dataset and refine the model's performance on labeled data of audio downstream tasks like speech recognition/translation and audio classification as shown in the figure 2.2

XLS-R demonstrates significant outperformance over the existing state-of-the-art models in speech processing tasks like recognition/translation/diarization and language identification, as documented in the official report **XLS-R**.

Jasper: Developed by NVIDIA AI researchers [51], Jasper is an automatic speech recognition (ASR) system rooted in deep learning principles. Despite its modest

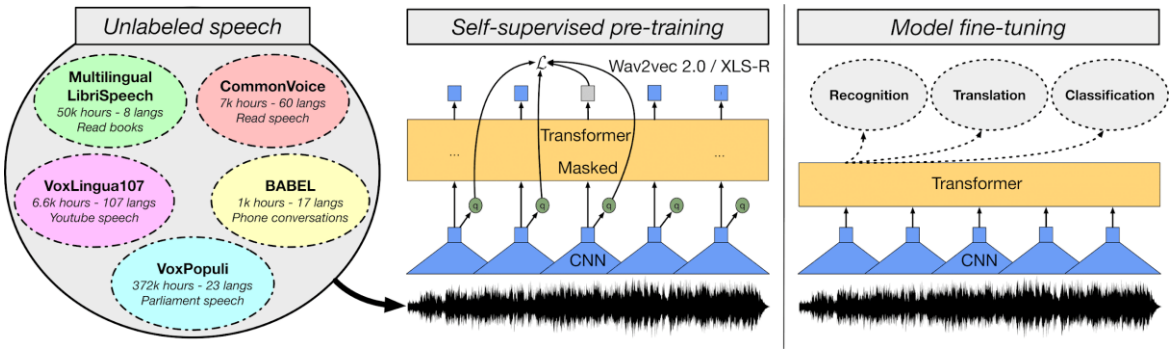


Figure 2.2: Wav2vec 2.0 / XLS-R

name, ”Jasper” (Just Another Speech Recognizer) excels in converting spoken language into textual format with high efficiency.

A distinguishing feature of Jasper lies in its utilization of convolutional neural net- works (CNNs) within its architecture. While recurrent neural networks (RNNs) have traditionally dominated ASR tasks, Jasper’s adoption of CNNs enables paral- lel computation, facilitating faster audio data processing and potentially reducing computational overhead compared to conventional RNN-based ASR models.

Jasper has demonstrated competitive performance across diverse ASR benchmarks, underscoring its capability to accurately transcribe spoken language. Its versatile architecture supports adaptation to various languages and acoustic environments, enhancing its applicability across a broad spectrum of ASR applications.

To conclude, Jasper represents a significant advancement in the pursuit of efficient and precise ASR models. Recognized for its robust performance, Jasper contributes to ongoing advancements in speech recognition technology.

2.0.2.2 Conformer:

Conformer, as described by Gulati et al. constitutes a deep learning framework tailored for automatic speech recognition (ASR) and natural language processing (NLP) tasks. It excels particularly in converting spoken language into textual form for ASR applications. Conformer architectures have garnered attention for their

adeptness in handling sequential data and achieving impressive performance across various ASR benchmarks.

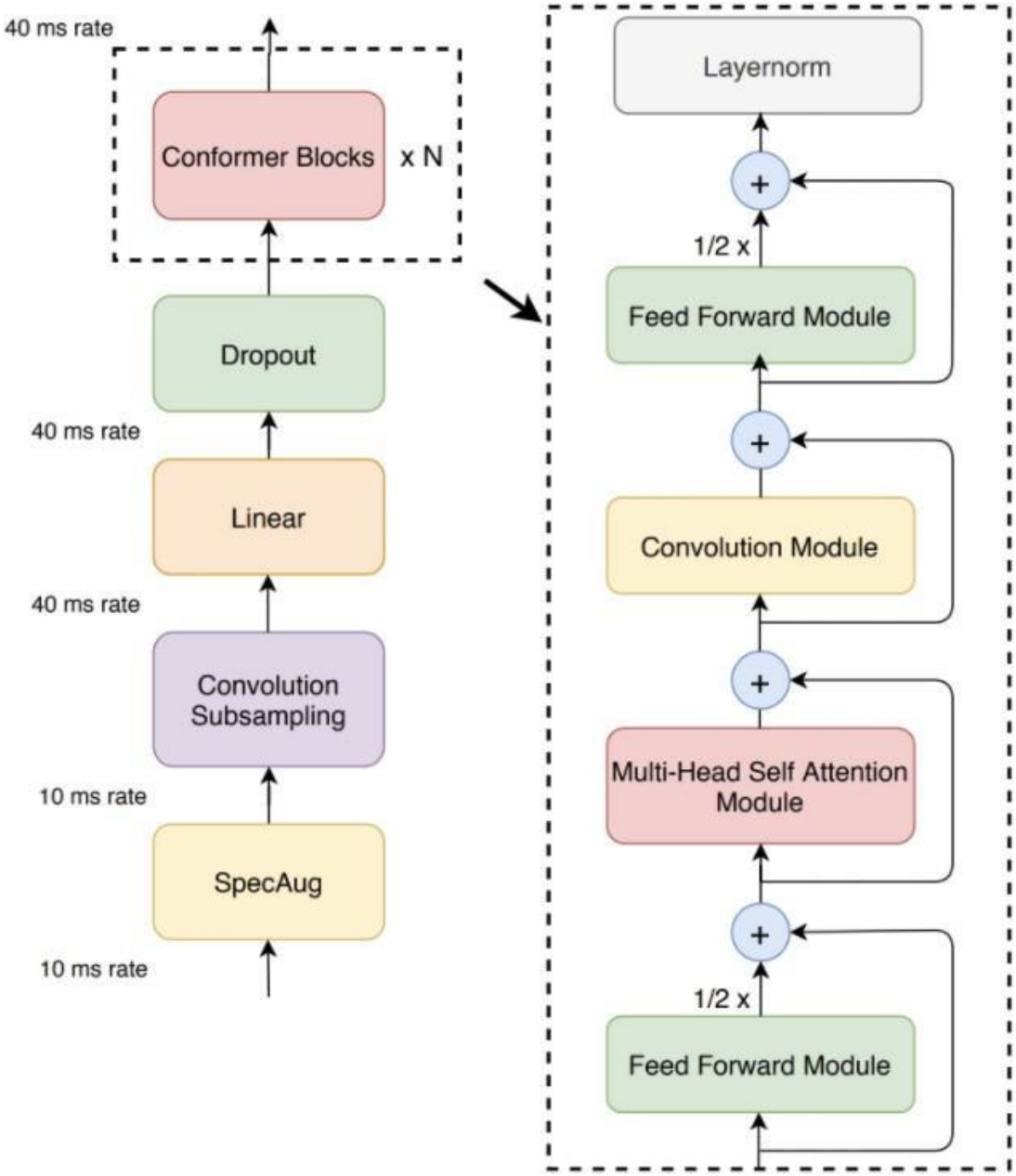


Figure 2.3: Conformer encoder model architecture

As depicted in Figure 2.3, the Conformer framework integrates several key com- ponents, including convolutional layers, a self-attention mechanism, and additional elements inspired by the Transformer architecture. The Conformer architecture con- sists of two macaron-like feed-forward layers that employ half-step residual connection

which encapsulate both multi-headed self-attention and convolution modules, followed by a post-layer normalization step.

The audio encoder first processes the input through a convolutional subsampling layer, subsequently passing it through multiple Conformer blocks, as illustrated in Figure 2.3. A distinctive feature of this model is the replacement of traditional Transformer blocks with Conformer blocks. Each Conformer block is structured with four stacked modules: a feed-forward module, a self-attention module, a con- volution module, and a final feed-forward module.

This hybrid approach equips Conformer models with the capability to capture local and global dependencies within input audio data, thereby enhancing their efficacy in interpreting spoken language.

Key features of Conformer models include:

1. **Parallelization:** Conformer architectures are optimized for high paralleliz- ability, facilitating accelerated training and inference processes. This attribute significantly enhances their efficiency in processing audio data.

2. **Self-Attention:** Leveraging a Transformer-inspired self-attention mechanism, Conformer models excel in capturing contextual dependencies across various segments of input sequences. This capability markedly improves their accu- racy in speech recognition tasks.

3. **Depth and Stacking:** Conformer structures typically employ multiple layers stacked sequentially. This design enables them to effectively capture intricate patterns and features present in audio data, enhancing their robustness and performance.

4. **Adaptability:** By training on appropriate datasets, Conformer models can adapt seamlessly to diverse languages and dialects. This adaptability renders them versatile for a wide spectrum of ASR applications.

Conformer-based ASR models have consistently achieved state-of-the-art perfor- mance across numerous ASR benchmarks, underscoring their efficacy in compre- hending and transcribing spoken language. Their adeptness in handling sequential data and their flexibility to accommodate various linguistic and environmental con- ditions have positioned them as a preferred choice for both academic research and

industrial applications in ASR.

2.0.2.3 Google ASR:

Developed by Google, Google Automatic Speech Recognition (Google ASR) [5] is a strong speech recognition system essential to many Google products and services including Google Assistant, Google Search, and Google Translate. This technology lets users vocally interact with their devices, where Google ASR precisely converts spoken input into a textual format for next analysis or use. Google ASR has these main traits and features:

1. **Accuracy:** Google Automatic Speech Recognition (Google ASR) is renowned for its high precision in converting spoken language into text. It employs advanced machine learning algorithms and extensive datasets to continually enhance its recognition accuracy.

2. **Multilingual Support:** Google ASR lets users interact with Google services in their preferred language by providing strong support for a varied global ar-ray of languages and dialects.

3. **Voice Commands:** Google ASR drives voice-command functionalities, em-powers users to manage devices, conduct web searches, schedule tasks, and execute other activities using spoken instructions.

4. **Voice Search:** Facilitating voice-based web searches, Google ASR enables convenient access to information and online resources without the need for manual typing.

5. **Accessibility:** Google ASR plays a pivotal role in accessibility initiatives, allowing individuals with disabilities to engage with technology through voice interactions.

6. **Natural Language Processing:** Integrated with sophisticated natural lan- guage processing (NLP) capabilities, Google ASR interprets spoken language in a conversational and context-aware manner, enhancing user interaction experiences.

7. **Cloud-Based Service:** Google ASR operates as a cloud-based service, offering developers the flexibility to integrate speech recognition functionalities into their applications and services.

8. **Privacy Considerations:** Google prioritizes user privacy and data security in its ASR implementations, ensuring users have control over their voice data and settings.

Google ASR continues to advance, fostering seamless and accessible voice interactions across a global user base. Its applications encompass a wide spectrum, encompassing voice assistants, search engines, transcription services, and voice-operated devices.

2.0.2.4 Whisper:

Currently garnering much attention, OpenAI's Whisper [19] is a discourse language model with great promise. Like other Whisper series models, Whisper-medium arrives pre-trained and is ready for use without additional fine-tuning required. For many uses, including speech-driven interfaces and transcription services, where striking a balance between speed and accuracy is crucial, this makes it a sensible choice.
Whisper has been painstakingly trained and developed, exposed to a varied dataset including almost 100 languages and over 680, 000 hours of well selected multilingual and multitask supervised data sourced from the web.

The model exists in two forms. English only and multilingual. The English-only models concentrate on speech recognition, in which case the spoken audio and the expected transcribed text are in the same language. Conversely, multilingual models are taught to translate as well as to recognize speech. These models in speech translation forecast text transcriptions in a different language than the spoken audio. It comes in many flavors. Table 2.2 shows the variances and specifics. Like Wav2Vec 2.0, it did fairly well on standard Bengali but quite badly on regional data. This is so even though it was trained on a large-scale dataset created from Bengali contents taken from the internet, it lacked even 1.5 hours of audio data. Especially, the training corpus does not have significant representation of

Table 2.2: Whisper model types details

Size	Parameters	English-only model	Multilingual model
tiny	39 M	Yes	Yes
base	74 M	Yes	Yes
small	244 M	Yes	Yes
medium	769 M	Yes	Yes
large	1,550 M	No	Yes

spontaneous Bengali speech with distinct regional accents and dialects, a gap addressed by the dataset introduced in this research.

As depicted in Figure 2.4, The Whisper model architecture, developed by OpenAI for automatic speech recognition (ASR) and language translation, is structured around a transformer-based encoder-decoder framework. The following are its key components:

1. **Encoder-Decoder Structure:** The architecture comprises an encoder and a decoder, facilitating efficient processing of audio input and the generation of text output.

2. **Audio Input Processing:** Raw audio signals are initially transformed into spectrograms, representing the frequency content over time, thus enabling structured analysis of the audio data.

3. **Multi-Head Self-Attention:** The model employs multi-head self-attention mechanisms, allowing it to attend to multiple segments of the input audio simultaneously. This capability is crucial for capturing long-range dependencies and contextual information.

4. **Positional Encoding:** Positional encodings are incorporated to maintain the sequential characteristics of audio data, providing the model with information regarding the temporal order of input frames.

5. **Feed-Forward Layers:** After the self-attention layers, feed-forward networks are utilized to further process the output, applying non-linear transformations that enhance the model's representational capacity.

6. **Layer Normalization:** Throughout the model, layer normalization is implemented to stabilize training and improve convergence rates.

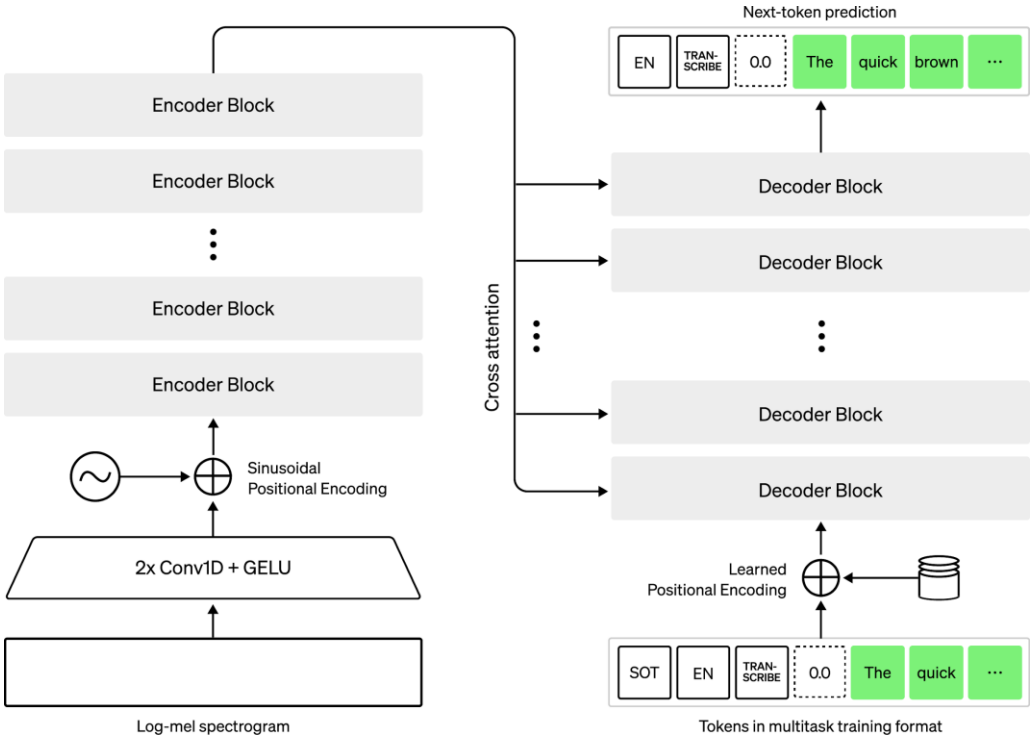


Figure 2.4: Conformer encoder model architecture

7. **Output Generation:** The decoder generates text sequences based on the encoded audio representation, utilizing both the attention mechanism and previously generated tokens to inform its predictions.

8. **Training Data:** The model is trained on a diverse and extensive dataset of multilingual speech, which significantly enhances its ability to handle various languages and accents.

All things considered, the Whisper architecture specifically addresses the difficulties with speech recognition and translation tasks while leveraging the strengths of transformer net-works.

This work especially addresses enhancing the training dataset with regional Bengali accents and dialects, so supporting current research projects. The dataset under discussion emphasizes the several advantages of using a thorough and varied dataset, so improving the robustness of the model in negotiating various accents, reducing background noise interference, and knowing specialized technical requirements.

including regional languages including Chittagong.

Beyond its contributions to speech recognition capabilities, Whisper's adaptability spans transcription and translation chores across many languages and dialects, transcending linguistic barriers and supporting cross-cultural communication.

Furthermore, Whisper's capacity to speed up the conversion of spoken materials from many languages into English promotes better intercultural communication and so increases access and understanding in many linguistic environments.

Looking ahead, deep learning algorithm developments promise ongoing increases in voice recognition accuracy, especially in managing variations including regional accents and dialects.

Chapter 3 Background Study

In this chapter, we will address the challenges associated with ASR modeling for regional dialects and provide a comprehensive linguistic analysis of each regional dialect.

3.1 Challenges with Bengali Speech Recognition with Regional Dialects

Among the people living in the Bengal region, Bengali is the most often used language. Though constant efforts toward standardization, regional dialects often prevail in daily communication instead of the standardized Bengali form. Not unique to Bangladesh, this trend reflects similar trends seen in other multilingual countries. Local dialects or languages are often used in daily contacts here, in contrast to the use of standardized language set aside for official documentation, media, and education. Like many languages, spoken language variance greatly from the accepted standard in daily settings.

3.1.1 Deviation Challenges

The standard form of Bengali encounters several challenges, including the preservation of cultural identity. There is worry as languages change that the process of standardizing Bengali might minimize some cultural quirks and regional traits unique in different dialects and local variants. For some demographic groups, accessibility and inclusivity of the official language provide difficulties. especially those with less education, which influences their capacity for comprehension and efficient use. This difference can cause social exclusion and impede access to knowledge, education, and possibilities. Moreover, proficiency in the standard language may confer advantages in education, employment,

and other domains, thereby exacerbating linguistic divides within society.

Another challenge involves the lack of modern terminology. Like all languages, Bengali grapples with the adaptation of new technological, scientific, and academic advancements. Developing and integrating new terminology while maintaining linguistic integrity presents a complex undertaking.

3.1.1.1 Regional Diversity

The Bengali speaking world has a wide range of so-called regional variation, accents, and idiomatic expressions due to language contact, culture, history and geography with neighboring regions. Most of these regional or local dialects are used in colloquial, home and informal situations. Such variation may be phonological, lexical or syntactic in nature, and may sometimes be markedly different from the standard Bengali in any or all of these features.

The relative natural progression also recognises that the usage and application of language are more fluid and adjustable over time and according to geographical location. However, dialectal variations provide a colorful and diverse form of Bengali which is actually the strength of Bengali speaking people as Bengali is a standardised means of communication for Bengalis belonging to all strata.

3.1.1.2 Cultural Diversity

In most cases, media outlets ensure that they stick to the standard language so as to avoid making loss in meaning or difference in pronunciation due to differences that are felt in different regions and sometimes may cause misunderstandings. Some categories of creative work, such as poetry, literature and other texts which often derive authority from creative rebellions against linguistic rules may suffer from a strict adherence to language standards, which this paper will argue.

3.1.2 Dataset Acquisition Challenges

The SUBAK.KO speech corpus [78] stands out as the singular reported speech cor- pus to date, involving participation from 61 native speakers representing all eight divisions and thirty-four districts of Bangladesh. But the recordings were uniformly in standard Bangla and were executed under controlled conditions within a studio environment.

Research and development in this domain have been hindered by insufficient datasets and resources.

3.1.3 Modeling Challenges

There are big computational challenges if researchers deploy SOTA models for training, especially if they are scholars without convenient access to such means. One must note that whereas Bengali language struggles to achieve this convergence during training several folds, comparatively easier than via English, primarily attributed to a larger number of unique tokens. Bengali is a typical agglutinative language and has various morphological differentiation in contrast with other languages; they are phonetically similar but have significantly different written script.

Furthermore, there are several varieties of Bengali regionalisms and variations in intonation, which poses even greater problems for linguistic modelling. While, as in the previous case, it is hard yet to determine the variability of phonetic features of these various dialects, addressing the issue of the linguistic diversity is crucial in the modelling process,

which explains why every model is the best at what they do. One of the key factors of our approach is the detailed information gathering process that captures a variety of linguistic nuances in both audio and text modalities. To provide neural networks with a solid dataset for addressing the complexity of ASR developments in Bengali, this careful curation's main purpose is to ensure that the training corpus carries different linguistic traits.

3.2 Available SOTA ASR Models

Bengali corpora has been used to refine many models, but datasets that are specifically devoted to Bengali speech that incorporate regional dialects are nonexistent. For this reason, none of these models have been refined in this context. A number of these models were used to compare performance to the corpus we created. The sections that follow give a thorough explanation of these models.

3.2.1 Google ASR

One of the key components of a number of Google services, including Google Assistant, Google Search, and Google Translate, is Google Automatic Speech Recognition (Google ASR) – an innovative speech recognition system developed by Google. This technology transcribes spoken language into writing for a number of purposes making the user interface understandable for the devices. Google ASR has evolved with time enabling users around the globe to have easier and seamless voice commands. Technologies like voice assistants, search engines, transcribe services, voice devices among others are just but a few of the areas where it can be applied. By and all, Google ASR is a useful tool that improves spoken language user's communication with Google services and products significantly.

For the purpose of this study, the test audio samples selected from the dataset used here were transcribed using the Google ASR API.

Google ASR overall performed very well for standard Bengali words, however when it was trained on different regional accents of Bengali its accuracy was very very poor.

3.2.2 Wav2Vec 2.0 Large

As stated in the official documentation XLS-R and they also show in Table 4, XLS-R has made substantial improvements over the prior art in many applications such as speech recognition, translation, diarization, and even language identification, while YellowKing[20] was a fine-tuned version of Wav2Vec 2.0 Large XLS-R.
For this purpose, we made inferences on the samples from the test data corpus created in this paper using our fine-tuned Wav2Vec 2.0 Large XLS-R model which was fine-tuned by Auditi Das and is available on the Hugging Face website [27]. The Common Voice 11.0 dataset [5] was used for fine-tuning this model.
When comparing the definite and variable characteristics of all the configurations and numerous variations that were tried, the specific version was called "Yellowking"

The lowKing was identified as the most stable and robust form of the King out of all the forms analyzed. This Wav2Vec 2.0 model was then tested, using the full test split to determine the results of this model.
While reasonable performance was seen when the model was tested on the standard

Bengali, its performance dramatically decreased when the regional speech data set was used across all the districts. The reduction in the model accuracy for words spoken in this region can therefore be expounded to the model's lack of training when it comes to handling dialects, accents or OOV words.

3.2.3 Tugstugi (Whisper-Medium)

There are many languages that are supported by OpenAI's Whisper-medium [19]; among these, Bengali language is also included in this transcription and translation linguistic or language model. Transcription of the spoken words in different language environments is not an easy task, but Whisper-medium fine-tuned from 680000hrs of transcribed speech data performs well in monolingual and multilingual environments. Thanks to the cross compatibility with two platforms it is possible to carry out the speech to text transcription and translation to improve language communication.

In our work, we have used an improved version of Whisper Medium , called Tugstugi [5] which is hosted in hugging faceModels. In short, Tugstugi stands out as the first winner in the Bengali AI Speech Recognition Competition. We have used the above model to test samples from the developed corpus on our developed tests. Moreover, two specific speechsets, such as the OOD-Speech, was utilized for building a big scale Bengali speech recognition dataset for out of distribution performances [24] that fine tune the model.

3.2.4 Hishab-Conformer

The Conformer is a deep learning framework designed specifically for ASR and NLP tasks in particular. Conformer models have garnered much attention due to excellent results in several ASR benchmarks and effectiveness in processing sequential data. Nevertheless, to arrive at more accurate outcomes, a self-attention layer and further modules borrowed from the Transformer are used in the proposed Conformer system, along with convolutional layers. Not only does this hybrid approach allow Conformer models to inherit capabilities of operating on sequential data,

Table 3.15: Fine-tuned Conformer Parameters

Parameters	Value
epochs	16
batch size	32
sampling rate	16kHz
use start end token	True
pin memory	True
number of workers	48
trim silence	False
max duration	18.5
min duration	0.2

ensures the use of their flexibility in the multiple lingual and environmental conditions which are why they are used in the scholarly and numerical ASR implementations. That conformer-based ASR models are robust and sufficiently able to capture spoken language has been shown through ASR evaluations that show state-of-the-art performance.

In our study, we compared our results with an improved FastConformer model [23] developed by Hishab. This model has been trained using the MegaBNSpeech corpus accumulated by the authors to 19,000 hours. Further experiments of the model were conducted using 4,000 hours of transcriptions from YouTube videos and the model was developed using NeMo Toolkit with the Conformer-CTC configuration. Subsequently, for the transcriptions of the train split, a byte-pair encoding tokenizer was first built [59].

Specifically, during the training, the pre-trained weights of Nemo English ASR were adopted for the initialization of the encoder weight. The training parameters have been provided in table 3.15 below.

Chapter 4

Methodology

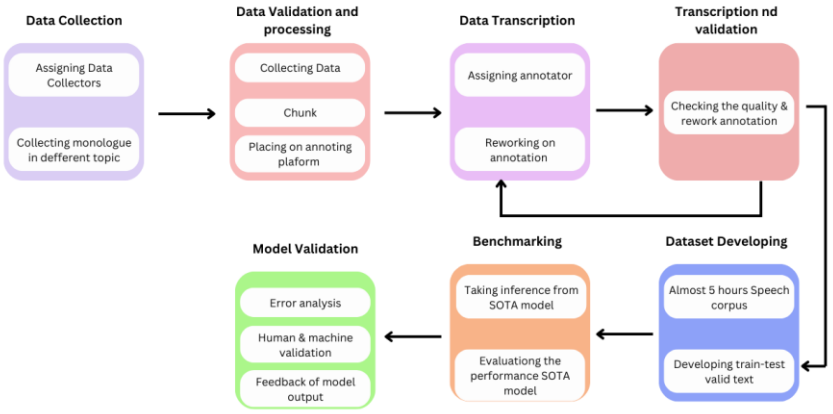


Figure 4.1: Top-level overview of the proposed methodology

For this current study, we created a speech corpus that mainly included regional dialects of Bengali and used an STT system which was created to transcribe spoken Bengali with various regional dialectical differences in that language as per a standard phonemic transcription norms set by linguists. The pipeline shown in Figure 4.1 presents an omnibus sequential workflow to accomplish the stated aim of the study. This is used as an efficient model for least-resourced languages hence hastening automating the STT progression for languages with many dialects.

and use different and sufficiently large variations and substantial linguistic differences.

The mentioned above diagram shows the detailed framework of the work to be done for our project. Since there are no English speech corpus freely available for Bengali language particularly for regional variety of accents and dialects in the public domain, we have taken the responsibility to build the corpus on our own. This process entails close observation of the patients in accordance with the guidelines that have been laid down that would help patients to speak as and when they wanted to. Also, through responding to thoughts and sentiment provocation, a compilation of monologues was achieved from people who expressed their points of view with no restrictive self- consciousness.

Based on the audio data validation and processing step, the recruitment of accurate and skilled transcribers for a diverse range of geographical areas that are in relation to the collected data was done deliberately. This decision was taken to maintain and further improve the quality of transcription thereby fully utilizing the collective experience of the transcribers with accent and dialect specific information content in speech data. For this reason, an extensive transcription test that included both the Ben-gali language as well as the regional dialects was conducted.

To enhance the quality assurance moreover, a linguist was hired and assigned to pronounce the credibility of the transcriptions and come up with remarks. The transcribers were then able to incorporate these suggestions and enhancements into the text and concurrent quality of the transcribed materials was bolstered.

Following the aforementioned processes, we curated the dataset and employed an 80:10:10 partitioning strategy in order to split it into training, testing, and validation sets. For further performance enhancement our model was fine tuned, using our in-house generated corpus, with a state-of-the-art (SOTA) model. Some qualitative benchmarking was also performed to determine the effectiveness and competitiveness of the proposed model, where the dataset was fed into other various SOTA models, and their transcription was compared.

Furthermore, we incorporated the design of an accent classifier to correctly segregate speech information using regional accents and/or dialects.

Currently, our dataset comprises nearly 5 hours of speech data from one distinct area within Bangladesh Dhaka: These places include Old Dhaka(Puran Dhaka) Hazaribagh, Lalbagh, Bongshal, Suritola, Tati-Bazar .

To summarize, through this work, we accomplished the following:

- Compiled nearly 5 hours of speech corpora specifically tailored to regional accents and dialects.
- Conducted both human and machine validation of the dataset.
- Generated benchmark results from various publicly available ASR models.

4.1 Data collection and validation

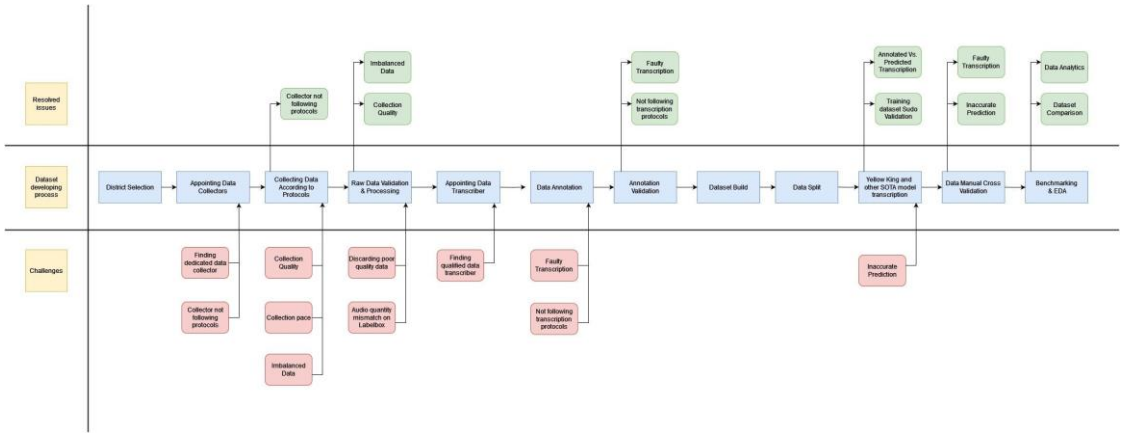


Figure 4.2: Detailed Workflow Diagram of Data Collection

Our objective was to develop a speech corpus that would enable any ASR model trained on it to accurately recognize Bengali speech from a variety of speakers across different topics, without exhibiting bias. To achieve this goal, several features were incorporated to ensure diversification, as outlined below.

1. Voice Diversification: To encompass a diverse array of voices and prevent bias towards any particular voice, we aimed to gather data from a wide range of individuals. Each audio clip was maintained at approximately ten minutes in duration. We targeted collecting at least 1 hour of data from each area.
2. Gender Diversification: To achieve a balanced gender representation, we aimed for a 50:50 ratio in the voice dataset. For a dataset comprising 5 hours of data, involving 30 individuals with each contributing 10 minutes of speech, we included 15 distinct male voices and 15 distinct female voices.
3. Age Diversification: Considering the changes in a person's voice with age, primarily due to reduced lung capacity and declining muscle strength and tone [26],

We incorporated age-related diversification into our speech corpus. This ensures that the ASR system trained on this corpus can effectively recognize the speech of speakers across various age groups.

4. Topic Diversification: To train an ASR system capable of recognizing speech across a wide range of topics and vocabulary, data collectors were guided to engage in conversations on everyday subjects. They were provided with a comprehensive list of topics, including sports, education, politics, family life, economics, and more. This approach aimed to ensure the system's proficiency in recognizing speech on virtually any topic.

5. Geographical Diversification: Geography significantly impacts accent adjustment [92], affecting various aspects of human interaction and linguistic diversity.

Incorporating these diversification features, we initiated data collection from various regions. The procedure is illustrated in Figure 4.2, with details discussed below.

4.1.1 Area Selection

Area were selected based on two criteria:

1. Available Acquaintances: Areas where personal connections existed were cho- sen to receive direct assistance from known individuals or to delegate tasks as needed.

4.1.2 Appointing Data Collectors

In-person visits were avoided to obtain spontaneous speech samples. Additionally, being non-native to the Area, there was a concern that speakers might not use their native accent while conversing with us and might not be as open compared to interactions with someone native to the area.

4.1.2.1 Challenges

1. Locating Committed Data Collectors: To ensure the one form of diversification it was crucial to employ a dedicated data collector who is rigorous to follow all the procedures of collection.

2. Adherence to Protocols by the Data Collector: As a result, it was necessary for the data collector to adhere strictly to the protocols in order to reach the intended diversification of the speech corpus. ASR models trained on this corpus may not meet expected results because of changes in data gathering.

4.1.3 Collecting Data According to Protocols

4.1.3.1 Resolutions

1. Collector Adherence to Protocols: Through manual validation of data collection, we could identify and correct potential data imbalances. This allowed us to guide the designated data collector in acquiring new data to ensure a balanced dataset.

4.1.3.2 Challenges

1. Collection Quality: The fact that the phone recorders were with the designated data collector also had some issues; the phone recorders give low volume since they are placed slightly away from the speaker's mouth; besides, there is an interfering noise in the recorded phone calls.

2. Collection Pace: Due to weather conditions, illness and other personal or business commitments the rate of data collection could vary and therefore reduce the reliability and comparability of the collected data.

3. Imbalanced Data: As will be later shown, there were some risks of dataset imbalance during the data collection process if the subjects were not closely supervised.

4.1.4 Raw Data Validation and Processing

4.1.4.1 Resolutions

1. Collection Quality: At this stage, we reviewed the audio clips, categorizing those with consistently low speaker volumes and identifying clips with significant back-

ground noise.

2. Imbalanced Data: During manual validation, we precisely identified the nature of data imbalances and devised methods to rectify them.

4.1.4.2 Challenges

1. After careful review of all gathered data, we deleted recordings displaying noticeably poor quality or that did not follow the specified protocols.

2. Audio Quantity Mismatch on Labelbox: We force-split any segments longer than 30 seconds after segmenting audio using the VAD technique so that every audio chunk fits this length limit. All the samples were then run through a filter meant to remove two particular kinds of chunks: those with sound activity but no speech and those with speech that is inaudible. We then single batch uploaded these samples to the Labelbox [94] data annotation tool. Sometimes a bug might stop all chunks from being uploaded at once without giving warnings. We had to run a script to confirm that every chunk had been effectively uploaded to Labelbox in order to solve this.

4.1.5 Appointing Data Transcribers

4.1.5.1 Challenges

1. Finding Dedicated Data Transcribers: We hired transcriber who were residents of the same area where the data was collected. This choice provided them with an advantage as their ears were accustomed to the local accents and dialects. Applicants for the data transcription position underwent assessments with the assistance of a linguist who evaluated their scripts. The linguist selected candidates who demonstrated a strong command of Bengali spelling, grammar, and proficiency in their native accent. Following a thorough evaluation process.

4.1.6 Data Annotation

I used the data annotation platform Labelbox to label the data. Transcribers were granted access to the data along with specific instructions and protocols provided by the linguist on handling the spellings of any out-of-vocabulary words. A screenshot of the interface of this annotation platform is shown in Figure 4.3.

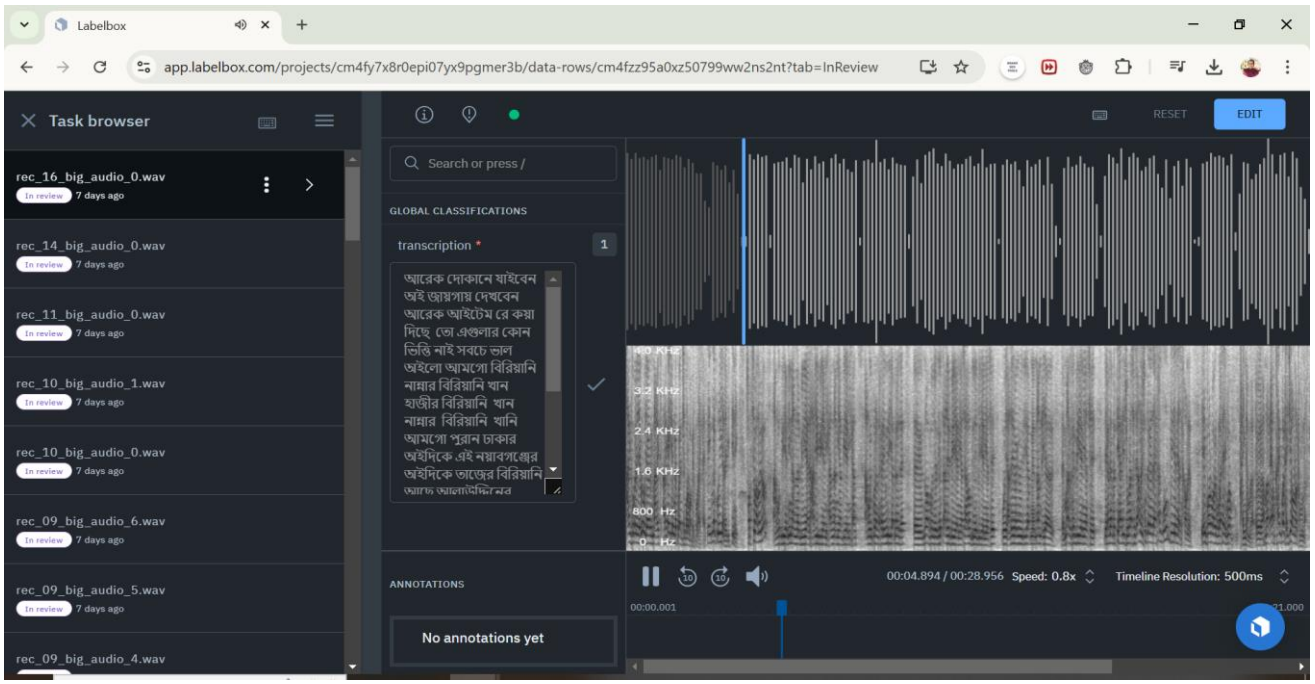


Figure 4.3: Interface of the annotation platform Labelbox

4.1.6.1 Challenges

1. Faulty Transcription: Despite clear instructions and adherence to established protocols, human annotators may still make errors during the transcription process, even if they are minor and unintentional.

2. Not Following Transcription Protocols: Certain individuals might become confused and fail to grasp the protocols, resulting in errors during the transcription process.

4.1.7 Annotation Validation

During the annotation process, a linguist diligently verifies spellings and transcriptions. Nonetheless, despite these efforts, occasional errors or inaccurate transcriptions may persist.

4.1.7.1 Resolutions

1. After all specifically completed transcriptions have been delivered to a CSV file with relevant data. Linguists then review, comment on this for a process of improvement of this file. Many transcribers subsequently edit the texts they have produced after considering the remarks made by the linguist regarding transcription errors. This last check confirms that the dataset is correct before its completion.

2. Not Following Transcription Protocols: If a transcriber finds themselves confused at some point or committed mistakes, they will correct them with the advice of the linguist or by using the feedback given by the said linguist.

4.1.8 Dataset Build

After completing all transcriptions of the audio files, we removed entries that were null or had other problematic issues. This process resulted in a dataset that includes all relevant information for each audio clip, along with their transcriptions.

4.1.9 Dataset Split

After building the dataset, We performed 80:10;10 train-test-valid splits of the entire dataset.

4.1.10 SOTA Model transcription inferences

In this section, we performed inference on all audio samples within our dataset using cutting-edge models such as Google ASR, Wav2Vec2 large, A fine-tuned FastConformer which the authors claimed to have trained using 20,000 hours of pseudo-labeled Bengali speech data named, YellowKing (Kaggle Competition winner on the Bengali Common Voice Speech Dataset), and Tugstugi (Kaggle Competition winner on the Out-of-Domain Speech dataset). Additionally, we developed a base model by fine-tuning a pre-trained model using our proprietary training dataset.

4.1.10.1 Resolutions

1. Annotated Vs. Predicted Transcription: We computed the Word Error Rate (WER) and Character Error Rate (CER) scores for each of these inferences and our own fine-tuned model by comparing them with the human annotations in the dataset.

2. Training dataset sudo validation: During manual validation conducted by a linguist, we identified data imbalances and implemented corrective measures.

4.1.10.2 Challenges

1. Inaccurate Predictions:

We assigned a linguist to cross-check and validate the model's predictions based on several criteria:

- Diffs.: Comparison of the model's predictions with the human annotator's transcription to assess substantial discrepancies.
- Word Error Rate (WER): Quantifies the percentage of incorrect words in the predicted text compared to the human transcript.
- Character Error Rate (CER): Measures the percentage of inaccurate characters in the predicted text.
- Word Insertion: Number of additional words not present in the human transcription.
- Word Deletion: Number of words missing from the human transcription.
- Word Insertion and Deletion Total: Combined count of inserted and deleted words by the models.

Based on these criteria, samples were categorized during the manual validation process into

- Incomplete Sentence: Instances where the ASR models failed to transcribe all spoken words in the audio clip.
- Incorrect Sentence: Cases where the ASR models transcribed incorrect words or included spelling mistakes in their transcriptions.

4.1.11 Data Manual Cross Validation:

4.1.11.1 Resolutions

1. Faulty Transcription: In this stage the number of errors in transcription is considerably low. The few other inconsistencies are detected at the final step that is, when the model's output is compared with the ground truth by the annotators, who quickly address them.
2. Inaccurate predictions: On reviewing the model predictions, identified issues are documented and addressed through appropriate solutions. After resolving these issues, the models are retrained to improve prediction accuracy.

4.1.12 Benchmarking and EDA

After obtaining all predictions, we proceeded to benchmark them against predictions made by other state-of-the-art (SOTA) models developed in previous steps.

4.1.12.1 Resolutions

Data Analysis: For the purpose of measuring and evaluating such parameters as the total accuracy of correct word and character predictions, more detailed data analysis was performed in this step. To be able to explain the results of the dataset, some visualisations were made as well.

Dataset Comparison: As we trained two different models for each datasets using the same parameters it gave us an opportunity to test our dataset against other datasets which were made available. We also evaluated and compared their predictions to determine if the obtained dataset was suitable for training models utilizing tasks specific for each instance.

This dataset was preprocessed and split for the modelling process before it was used to feed the model. These are the details of these steps:

4.2 Dataset Split

The entire dataset, totaling almost 5 hours, was divided into train, test, and validation splits with a ratio of 80:10:10. The dataset consisted of 450 samples, totaling approximately 5:03:55, 2:51:11, and 2:51:43 hours of data, respectively. Dur- ing this division, care was taken to ensure that recordings from a single speaker were included in only one of the splits. This approach facilitates a fair assessment of speaker generalization.

4.3 Proposed Methodology

The standard implementation of Wav2vec2.0, in conjunction with the CTC algorithm, does not require a language model head or dictionary for transcription decoding. Models

like Wav2vec2.0 utilize the Wav2Vec2 Tokenizer, which typically performs tokenization at the subword level as shown in the simplified architecture in Fig 4.4.

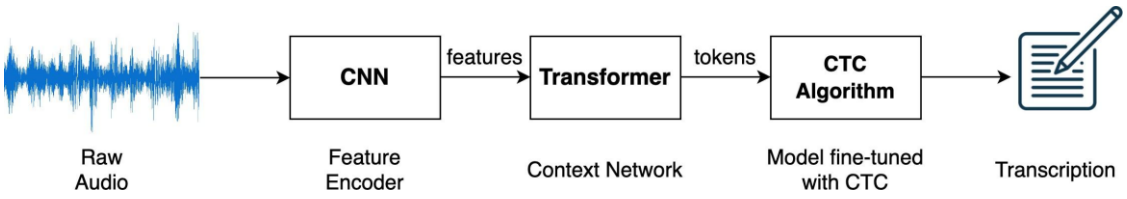


Figure 4.4: Simplified architecture of Wav2vec 2.0

This ASR system leverages self-supervised learning and transformers to efficiently decode audio signals into textual characters, making it particularly well-suited for low-resource languages like Bengali. The CNN encoders focus on feature extraction from the raw audio signal, while the masked modeling and transformer-based context captur- ing enable robust representations. The character gram language model further refines the output by modeling dependencies between characters, ensuring coherent transcriptions aligned with Bengali phonetics and grammar. A detailed breakdown of each component of this system is described below.

1. Raw Waveform (Input): The system begins by receiving the raw audio wave- form as input, representing the speech signal of the speaker. The waveform is a continuous-time signal containing rich acoustic information, which must be processed into discrete feature vectors for further analysis.

2. CNN Encoders (Feature Extraction): The raw waveform is passed through a se- ries of CNN layers, which act as feature extractors. During this phase, the model learns a latent representation of audio, encoding it into high-level features that are abstracted from the raw signal converted by the CNN encoders. These features capture important information from the input audio, like speech patterns, and reduce the temporal dimension of the input, enabling the extraction of essential features such as phonetic information and temporal dependencies and are passed to the fine-tuning phase. However, they do not explicitly map to acoustic units like phonemes or syllables.

   These feature vectors form the backbone of the speech encoding process, feeding into the subsequent layers for more abstract representation.

3. Masked Input Vectors (Self-Supervised Learning): In this stage, portions of the feature vectors are masked (hidden), indicated by the ”MSK” markers. This is an essential part of the self-supervised pre-training phase in models like wav2vec 2.0. The model is tasked with predicting the masked portions of the input, forcing it to learn robust and informative representations of the speech signal without rely- ing on labeled data. This method allows for efficient learning from large amounts of unannotated speech data, which is particularly beneficial for low-resource languages like Bengali.

4. Transformer Encoder Block (Contextual Representation): The masked input vec- tors are fed into a Transformer Encoder Block. The transformer architecture, fea- turing Multi-Head Attention mechanisms, enables the model to capture long-range dependencies within the speech sequence. It effectively models the global context by attending to different parts of the input sequence simultaneously.

   This is crucial in ASR systems as speech sounds can have contextual dependencies that are far apart in time. By employing multiple attention heads, the transformer encoder can focus on various parts of the input sequence and integrate contextual information from diverse perspectives.

   The Add and Normalize layers ensure that the model retains stable gradients during training and maintain layer consistency after the multi-head attention operations.

5. Encoded Token (Character Classes): The output from the transformer encoder

block is a sequence of context-rich encoded tokens. These tokens represent high-

level abstractions of the input speech signal and are mapped to a specific class, which in this case, corresponds to characters in the Bengali alphabet. The model tries to predict the next character based on the acoustic features, and decoding happens on a per-character basis, where each encoded token translates to a character from the transcript. Each token is representing learned latent speech representations, an embedding containing both local acoustic information and global context, suitable for decoding into text.

6. Character Gram Language Model (LM): A character-level n-gram language model is used to decode the sequence of encoded tokens into a sequence of characters. This model estimates the probability of a character sequence given the context. The equation shown represents the joint probability of the character sequence $C_1, C_2, \ldots, C_n$ being the correct transcription, which is computed as the product of individual conditional probabilities:

$$P(C1, C2, .. , Cn) = P(C1) \times P(C2 \mid C1) \times P(C3 \mid C1, C2) \times P(Cn \mid C1, C2, ., Cn-1) \quad (4.1)$$

   This ensures that the decoded output respects the character-level dependencies typical in Bengali regional scripts, improving the quality of transcription.

7. Predicted Outputs (Textual Representation): Finally, the system generates the transcription of the input speech and it is the predicted character sequence as a text. The model does not employ an acoustic unit discovery system; rather it takes the belief that audio features are mapped to characters as outputs. Here the original input sound is transcribed into a readable form in the target language putting into play the predicted characters and these substitute the Bengali regional dialect texts.

4.4 Benchmarking

Therefore, we have chosen several state-of-the-art ASR models for Bengali with which we compare our improved model. We leveraged the commonly employed speech APIs and the benchmark models commonly employed by the community. Here we employ the dataset we carefully chose for a range of benchmarking. The widely used speech API we chose is Google's speech-to-text cloud service API based on the Conformer model.

As for the STT, Google's cloud STT API was used in its unaltered form and setup. On the other hand, Meta's Wav2Vec 2.0 was benchmarked using a more improved Wav2Vec

2.0 model which [5] has been trained using Bengali Common Voice Speech Dataset [5]. Additionally, we benchmarked two competition winner models: Tugstugi [5], the Bengali winner, and Yellowking [20], the DLSprint Competition winner.Bengali-speech AI Speech Recognition Competition. Tugstugi is a fine tuned model derived using Whisper Medium [19] and Yellowking is a Wav2vec2 large [56] model optimized from two different speech datasets. Tugstugi was refined on OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking while, Yellowking was trained on the Bengali Common Voices Dataset [16] for about 70 hours Kaggle GPUs on and in Bengali using Speech & Speech recognition corpus [24]. We also evaluated the Hishab Conformer [23] model which the authors asserted was trained on 20K hours of Bengali speech data which was pseudo-labeled.

Based on the samples selected from the test split of the dataset created specifically for this work, we collected inference results of all standard and available deep learning models for ASR.

To assess how well our proposed model is performing as compared to the other models, we obtained the average Word Error Rate (WER) and Character Error Rate (CER) of all the inf cities of all the models with regional human-annotated ground truth data. The outcomes are depicted in Table 6.2.

Chapter 5

Dataset Statistics, EDA and Feature Extractions

In this chapter, we presented all the statistics, exploratory data analysis, and pertinent feature extraction from our developed speech corpus with regional Bengali dialects.

5.1 Dataset

This section shows all the corpus statistics of our corpus. In this section, we also illustrated how this regional speech corpus deviates from any standard Bengali speech corpus. We have used OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking [87] as the reference corpus for standard Bengali.

5.1.1 Bengali Speech Corpus with regional dialects

The dataset encompasses one specific place in Bangladesh Dhaka: Old Dhaka (Puran Dhaka) such as Hazaribagh, Lalbagh, Bongshal. Predominantly comprising spontaneous speech, it also includes a small amount of monologues and phone-recorded conversations from each area. Further details specific to each district are provided in the following table 5.1.

Table 5.1: Overview of Bengali Speech Corpus Regional Dialects (Puran Dhaka)

Duration	Total chunks	Average chunk size	Type
49 Minutes	70	19.55 seconds	Spontaneous
50 Minutes	73	20.71 seconds	Spontaneous
56 Minutes	90	23.58 seconds	Spontaneous
47 Minutes	54	20.67 seconds	Spontaneous
53 Minutes	91	18.51 seconds	Spontaneous
48 Minutes	72	20.51 seconds	
5 Hours 3 Minutes	450	21.604 seconds	Spontaneous

5.1.2 About the corpus

The corpus has 450 chunk samples resulting from 30 speech recordings where the data was split into 80:10:10 train-test-valid split resulting in 5 hour 3 minutes, 3 hour 51 minutes, and 3 hour 51 minutes of audio data in each fold. In the metadata, Each chunk is accompanied by an 'External_ID', 'Contents' which is the chunk transcription done by human annotators.

A detailed overview of the corpus is shown in the table 5.2

Table 5.2: Regional Speech Corpus Statistics.

‹→ denotes subsets | WPM = Avg. Words Per Minute | WPS = Avg. Words Per Sample | H:M:S = Hour(s) : Minute(s) : Second(s)

| OOV = Words Out of Canonical Standard Bengali Vocabulary in comparison to the unique words of the corpus

| Annotation Complexity is measured by the time needed to annotate every unit of data.

Breakdown of the dataset (Puran Dhaka)

subset	sample	Duratio H:M:S	Avg Rec Len	WPM	WPS	Uniq. Word	OOV%	Characteristic Phones Pair Count	Avg. Phone length perc(%)	Annot.Compl.
Dialect Dataset (Cumulative)	450	5:03:00	15.746	125.456	30.96	3250	68.6	8	0.982	110.5
Dialect Train	350	4:00:00	15.706	121.466	31.43	2860	60.2	-	0.986	70.87
Dialect Test	50	0:55:00	15.680	115.56	25.6	745	48.97	-	0.991	220.65
Dialect valid	50	0:55:00	15.670	114.52	24.01	710	48.10	-	0.997	221.34

5.1.3 Corpus statistics

The complete corpus consists of 450 audio segments, totaling 5 hours, 3 minutes, and 49 seconds of data collected from Puran Dhaka. The average segment length is 15.746 seconds, with a speech rate of 125.456 words per minute and an average of 30.96 words per segment. The corpus contains 3250 unique words, with 68.834% out of words vocabulary.

5.1.3.1 Train fold statistics

The train split comprises 350 audio segments, amounting to 4 hours of data collected from Puran Dhaka. The average segment duration is 15.706 seconds, with a speech rate of 121.466 words per minute and an average of 31.43 words per segment. The corpus includes 2860 unique words, with 60.2% out of vocabulary words.

5.1.3.2 Test fold statistics

The test split consists of 50 audio segments, totaling 55 minutes, and 11 seconds of data gathered from Puran Dhaka. The average segment duration is 15.680 seconds, with a speech rate of 115.56 words per minute and an average of 25 words per segment. The corpus contains 745 unique words, with 48.97% out of vocabulary words.

5.1.3.3 Valid fold statistics

The validation split comprises 50 audio segments, totaling 55 minutes, and 43 seconds of data collected from Puran Dhaka. The average segment duration is 15.*6*70 seconds, with a speech rate of 114.52 words per minute and an average of 24.01 words per segment. The corpus includes 710 unique words, with 48.10% out of vocabulary words.

5.1.4 Corpus Diversifications

5.1.5 Word and Grapheme Diversity

Figure 5.2 displays the different regions from which the data were collected, and table 5.3 illustrates the unique phonetic elements from each of those regions. In each phoneme pair, the first phoneme represents the conventional standard, while the second is the local variant used by inhabitants as a substitute for the standard phonetic sound.

Figure 5.2: Mapped regions of the dialects along with reference point

Geographical distance between regions makes linguistic diversity more noticeable. Conversely, geographically adjacent or closer-together regions typically show less graphemic diversity, indicating a higher likelihood of mutual intelligibility. The linguistic features of the Dhaka area, more especially **Puran Dhaka**, are different, exhibiting distinctive dialects that make it stand out. This particular group of dialects represents a localised linguistic identity and is exclusive to **Puran Dhaka**. For a better understanding, Table 5.4 shows how different regions use different word choices to articulate the same sentence.

Table 5.4: pronunciation of the sentence with IPA table

Bengali text	IPA Transcription
হ ভাই মোকলেছ্ চান অই মোটকা দুইডা আমগোরে দইরা নিয়া এমন মাইর দিছে	/ɦaɪ̯ bʱaɪ̯ mɔkleʃ tʃan ɔi̯ moʈka duɪ̯ɖa amgore ɖoɪ̯ra ni̯a ɛmon maɪ̯r ɖiʃe/
খাওন খাইতে আহো খাইতে আহো কি মিয়া হান্দায় না ভিত্তে কল্লা পর্যন্ত ডুবাইয়া	/kʰaon kʰaɪ̯te aɦo kʰaɪ̯te aɦo ki mi̯a ɦanɖaɪ̯ na bʱiʈʈe kɔlla pɔrtʃɔnto ɖubai̯a/
উইও অইদিন কা এই কামডাই করছে একান্ন টা ঘর লিয়া লিছে আড়াইশ	/ui̯o ɔi̯din ka ɛi̯ kamɖaɪ̯ kɔrʧe ɛkaɳnɔ ʈa gʱɔr li̯a liʧe aɽaɪ̯ʃo/
খাউটা হ অইটা অইটা খাউটা পিচ্চি বড় অয়না অইটার তে	/kʰaʊʈa ɦɔ ɔi̯ʈa ɔi̯ʈa kʰaʊʈa piʧ:i bɔɽɔ ɔi̯na ɔi̯ʈar ʈe/
কেয়ামত অইয়া গেলেও তুমি বিরানি খাইবা হ বাই তুমি তুমার আল্লার দোহায়	/kejamoʈ̲ ɔi̯a ge̯ʟeo ʈ̲umi birani kʰaɪ̯ba ɦɔ bʱaɪ̯ ʈ̲umi ʈ̲umar a̲ll̲ar ɖ̲oɦaɪ̯/

5.1.6 Voice or speech Diversity

Our goal when creating this corpus was to make sure that the speech data had as much diversity and representation as possible. We made an effort to standardise each sample's duration to roughly ten minutes. The data collector was told to collect more, shorter samples from other people to make up for any samples that lasted longer than ten minutes. On the other hand, the collector was instructed to get longer samples from other participants if the sample lasted less than ten minutes. Understanding that vocal traits change with age, we also made an effort to use our designated data collectors to gather speech data from people of all ages in each region.

5.1.7 Gender Diversity

This speech corpus comprises data from a minimum of 40 speakers, including approximately 80% male and 19.7% female participants. Additionally, 7.303% of the samples feature multiple speakers from both genders, resulting in 33 male speakers, 7 female speakers, and 5 clips with multiple speakers representing both genders.

5.1.8 Geographical Diversity

Although a single region may encompass multiple dialects [32], data collection was conducted across various subregions within each region to capture the diversity of dialects and ensure geographical representation. In total, 31 subregions were covered across

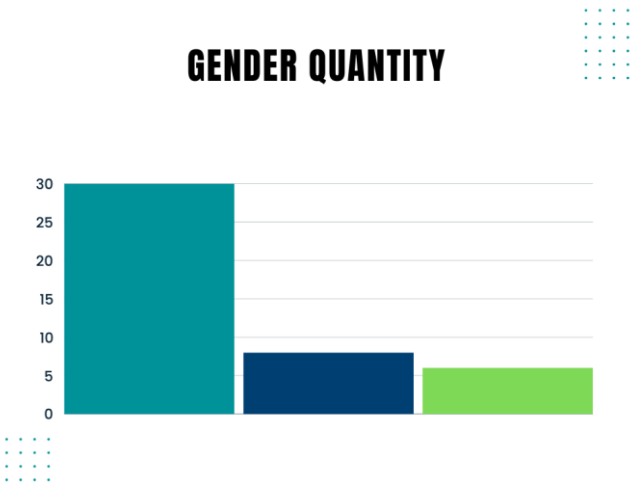


Figure 5.3: Gender quantity in the Regional Speech Corpus

Table 5.5: Data collected from each area

Area
Hazaribagh
Bongshal
Lalbagh
Suritola
Tatibazar

5.1.9 Topic diversification

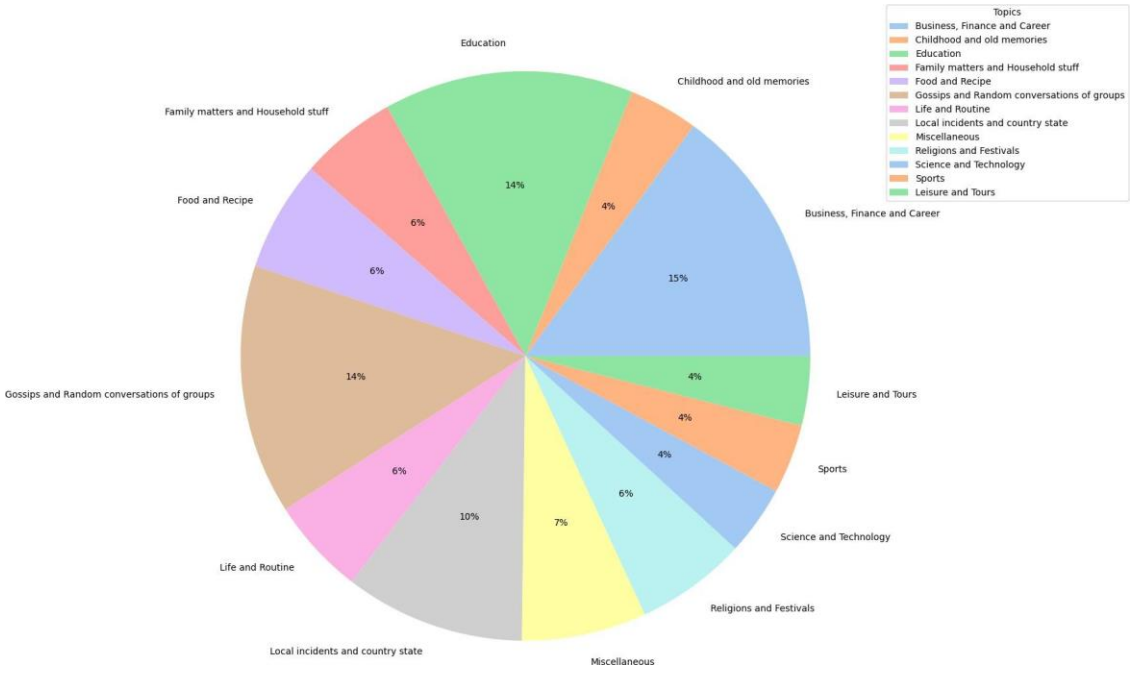


Figure 5.4: All subcategories of topics in the dataset

The dataset comprises speeches on 64 unique topics, categorized into 13 distinct categories. These categories are further organized into various subgroups of clusters, as illustrated in Figure 5.4.

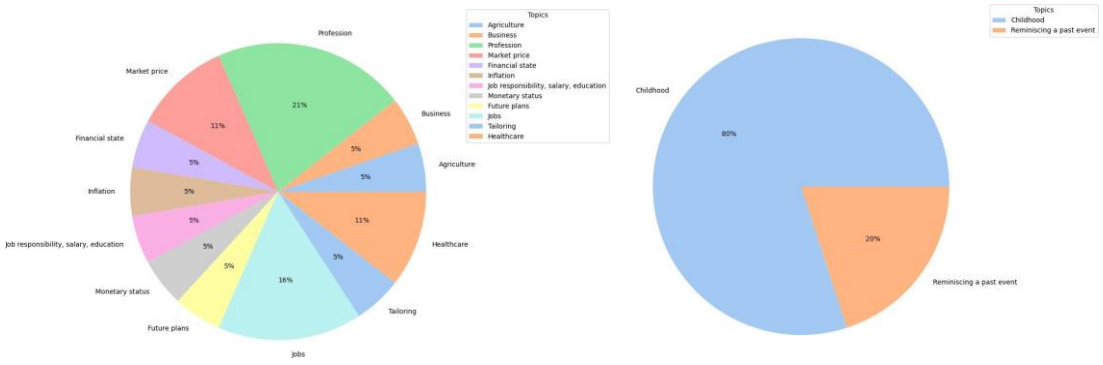


Figure 5.5: (Left) Topics in the ”Business, Finance and Career” subcategory and (Right) Topics in the ”Childhood and old memories” subcategory

5.1.9.1 Business, Finance and Career

This subcategory encompasses topics related to Business, Finance, and Career, as detailed

on the left of Figure 5.5. The majority of the topics pertain to individuals' professional

roles, job responsibilities, educational qualifications, and various professions through which they earn a livelihood. Given that a significant portion of the data was collected from rural areas, discussions frequently focused on agriculture, tailoring, and healthcare. Additionally, the subcategory includes topics such as market prices, inflation, financial conditions, and the monetary status of families or individuals.

5.1.9.2 Childhood and old memories

This subcategory encompasses topics related to nostalgia, as depicted on the right of Figure 5.5. It includes discussions about childhood experiences, past living conditions, and reflections on various events from the past. The majority of the conversations focus on childhood memories.

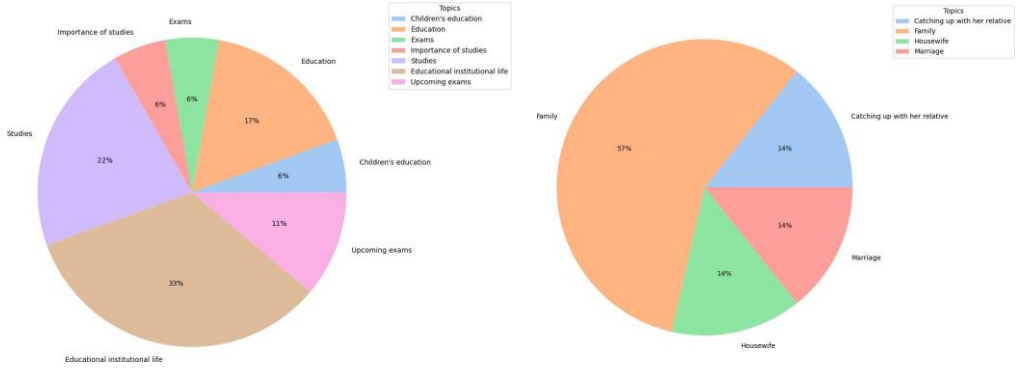


Figure 5.6: (Left) Topics in the ”Education” subcategory and (Right) Topics in the ”Family matters and Household stuff” subcategory

5.1.9.3 Education

This subcategory addresses topics related to education, as detailed on the left of Figure 5.6. It primarily features discussions by students about their examinations, academic institutions, and the subjects they are studying.

5.1.9.4 Family matters and Household stuff

This subcategory encompasses topics related to family and household affairs, as illustrated on the right of Figure 5.6. The discussions are predominantly conducted by older women and focus on familial matters and domestic issues within their households.

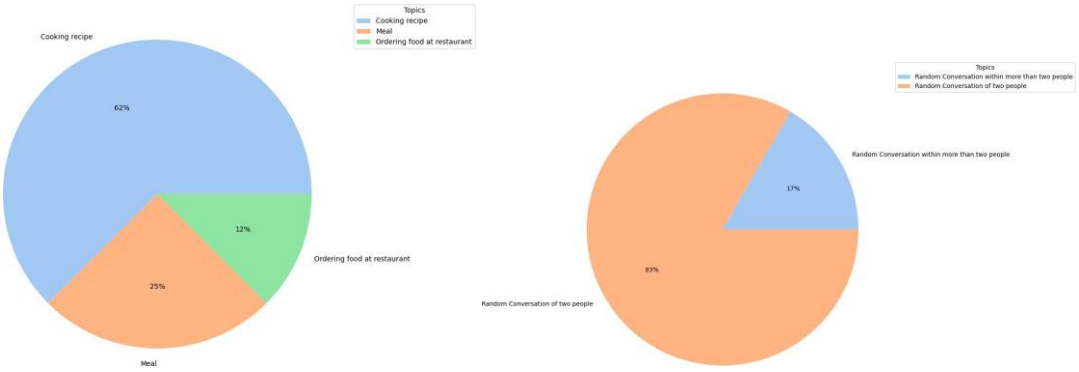


Figure 5.7: (Left) Topics in the ”Food and Recipe” subcategory and (Right) Topics in the ”Gossips and Random conversations of groups or individuals” subcategory

5.1.9.5 Food and Recipe

This subcategory addresses topics related to culinary matters, as depicted on the left of Figure 5.7. The discussions are primarily conducted by women and housewives, focusing on various recipes and food-related topics.

5.1.9.6 Gossips and Random Conversations of Groups or individuals

This subcategory primarily encompasses casual gossip or conversations among two or more individuals within a group, as illustrated on the right of Figure 5.7. The discussions often shift between topics, reflecting the spontaneous nature of the speech. The data collector was instructed not to interfere during these interactions. In some sample clips from this subcategory, there are more than two speakers, including the data collector.

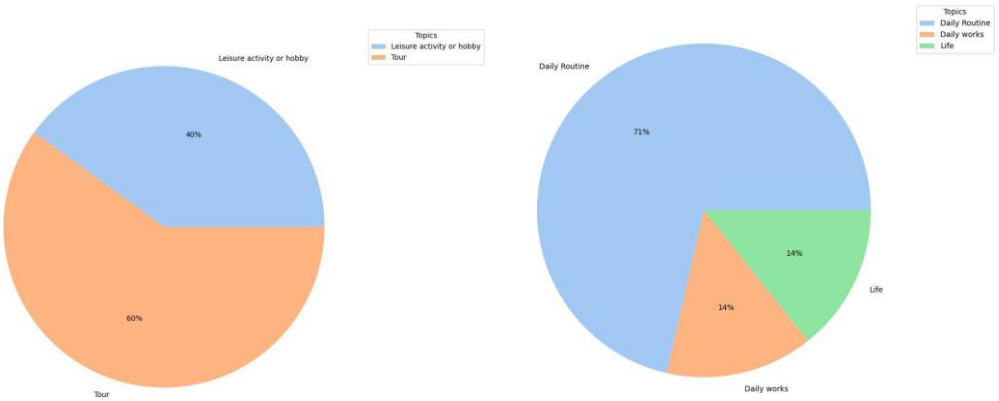


Figure 5.8: (Left) Topics in the ”Leisure and Tours” subcategory and (Right) Topics in the ”Life and Routine” subcategory

5.1.9.7 Leisure and Tours

This subcategory includes topics related to hobbies and travel, as shown on the left of Figure 5.8. The discussions cover various hobbies pursued by the speakers and the different trips they have planned or experienced with friends to various locations.

5.1.9.8 Life and Routine

This subcategory encompasses topics related to individuals' daily life and activities, as illustrated on the right of Figure 5.8. The majority of the speakers are students, and their discussions focus on the routine activities and daily experiences they encounter.

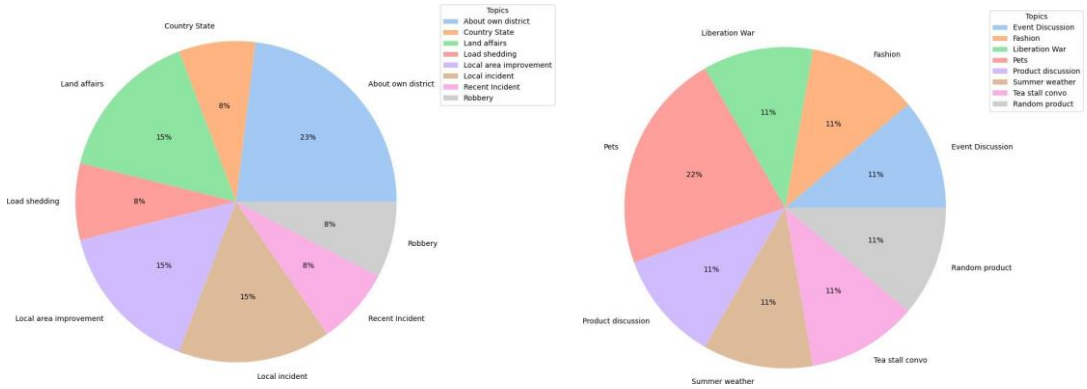


Figure 5.9: (Left) Topics in the "Local incidents and country state" subcategory and (Right) Topics in the "Miscellaneous" subcategory

5.1.9.9 Local incidents and country state

This subcategory addresses topics related to recent incidents occurring in the speakers' vicinity, as shown in Figure 5.9. The speakers, predominantly from rural areas, discuss various local issues, potential solutions, district-specific matters, local politics, and recent events within their communities.

5.1.9.10 Miscellaneous

This subcategory includes topics related to various miscellaneous subjects, as illustrated in Figure 5.9. The discussions encompass a range of topics, including different types of pets, fashion, and other diverse interests.

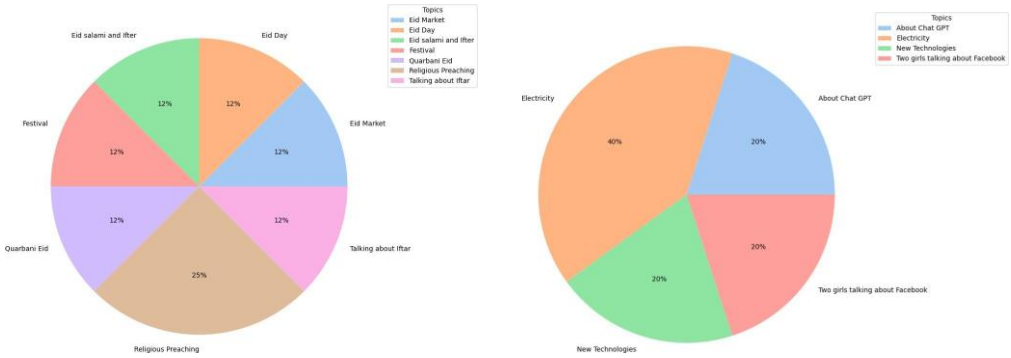


Figure 5.10: (Left) Topics in the ”Religions and Festivals” subcategory and (Right) Topics in the ”Science and Technology” subcategory

5.1.9.11 Religions and Festivals

This subcategory addresses topics related to religious activities. Given the predominantly Muslim population in the regions, the conversations primarily focus on Eid and associated practices. The topics and their distributions are presented on the left of Figure 5.10.

5.1.9.12 Science and Technology

This subcategory encompasses topics related to technology, as detailed on the right of Figure 5.10. Given that the majority of the data were collected from rural areas where stable electricity is a prevalent issue, discussions predominantly focus on this topic.

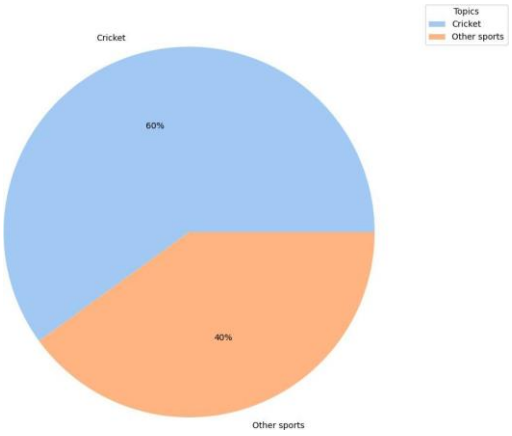


Figure 5.11: Topics in the ”Sports” subcategory

5.1.9.13 Sports

This subcategory includes topics related to sports, as shown in Figure 5.11. Given that cricket is the most popular sport in the subcontinent, the majority of discussions within this subcategory focus on cricket.

5.2 Exploratory Data Analysis and Feature Extraction

By conducting some exploratory data analysis and pertinent feature extraction, we explored our developed speech corpus with regional Bengali dialects in depth in this chapter. We also demonstrated how this regional speech corpus differs from any standard Bengali speech corpus. The reference corpus for standard Bengali is OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking [24].

5.2.1 Exploratory Data Analysis

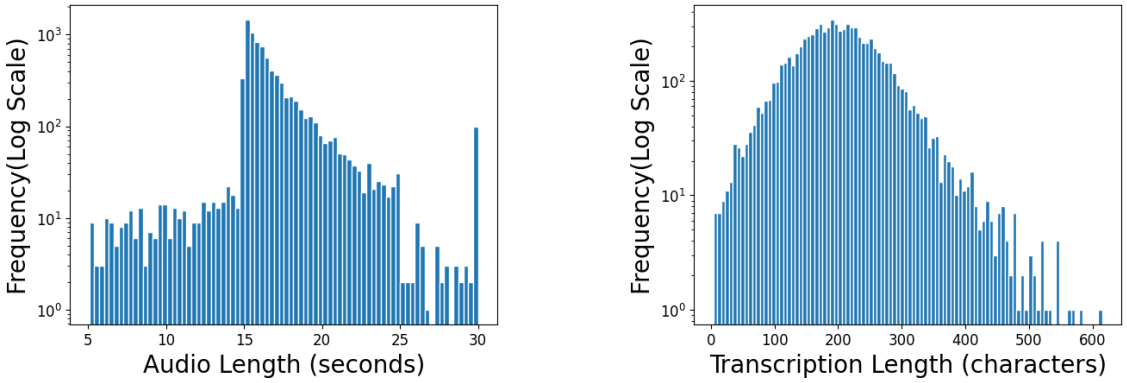


Figure 5.12: (Left) Audio length distribution of the regional Bengali dialect corpus (Right) Transcription length distribution of the regional Bengali dialect corpus

On the left of Figure 5.12 illustrates that the majority of the recordings have durations between 15 and 25 seconds, with a maximum length of 30 seconds. Additionally, on the right of Figure 5.12, it is demonstrated that there is no significant correlation between transcript character count and audio length.

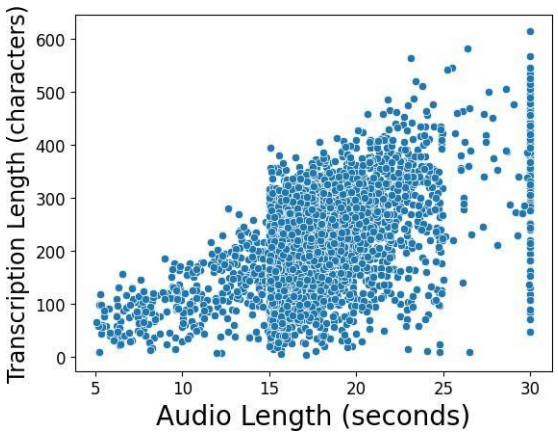


Figure 5.13: Transcription length vs audio length distribution of the regional Bengali dialect corpus

Although we do not observe any peculiar long transcripts for shorter audio recordings, the opposite is observed in several cases as we can see in Figure 5.13. Upon further investigation, these samples were found to be characterized by high levels of background noise, incomprehensible speech, or transcription errors.

5.2.2 Feature Extraction

5.2.2.1 Comparison with Standard Bengali

To evaluate the feature diversity within the Promito Bengali or Standard Bengali language dataset, we extracted Geneva speech features from 10, 000 samples from the referred standard Bengali speech corpus [54] and compared them with our regional Bengali dialect speech corpus.

We further examined the distribution shift using Figure 5.15, which focuses on a specific feature, with 'ancholic' encompassing all the regional Bengali (Puran Dhaka) dialect audios and 'Promito' representing 10, 000 samples from Standard Bengali speech data.

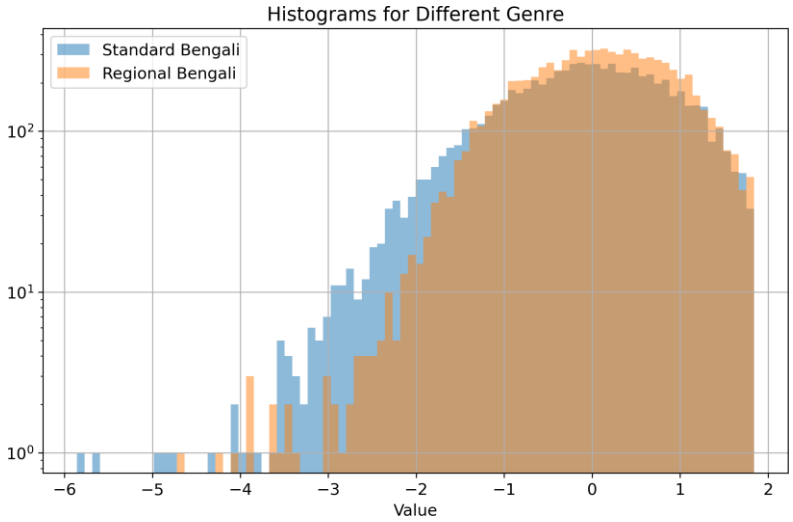


Figure 5.14: Histogram comparison between Geneva features of samples for Standard

Chapter 6

Result Analysis

6.1 Evaluation Criteria’s

We evaluated the model based on two evaluation criteria, Word Error Rate (WER) and Character Error Rate (CER).
The final WER and CER of this model are respectively Avg 0.728 and 0.347. The table 6.1 shows average WER and CER from three model’s results.

Table 6.1: Avg word error rate and character error rate

Word Error Rate (WER)	Character Error Rate (CER)
0.728	0.347

6.2 Model Inferences

Some region-wise inferences of the samples from our fine-tuned model are shown below in the figures 6.2, 6.3, 6.4 for the respective test data from the corpus.

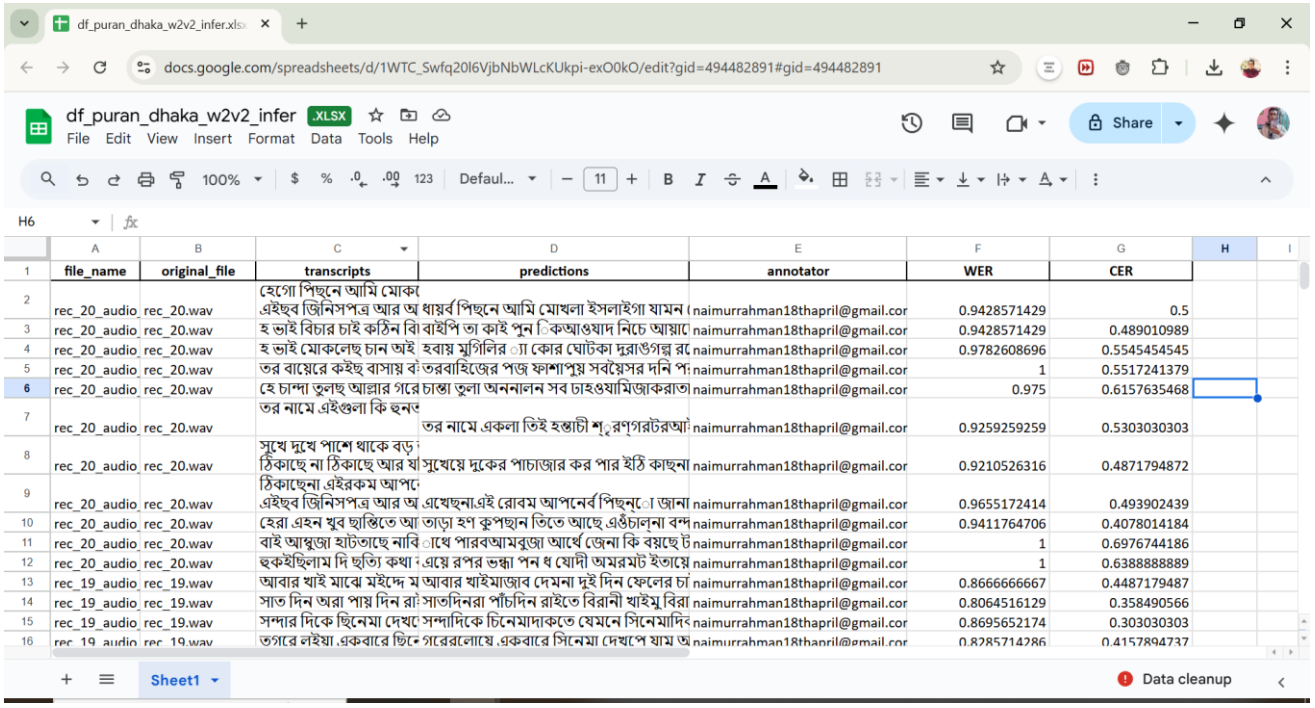


Figure 6.2: Model inferences samples on Puran Dhaka by Wav2Vec2

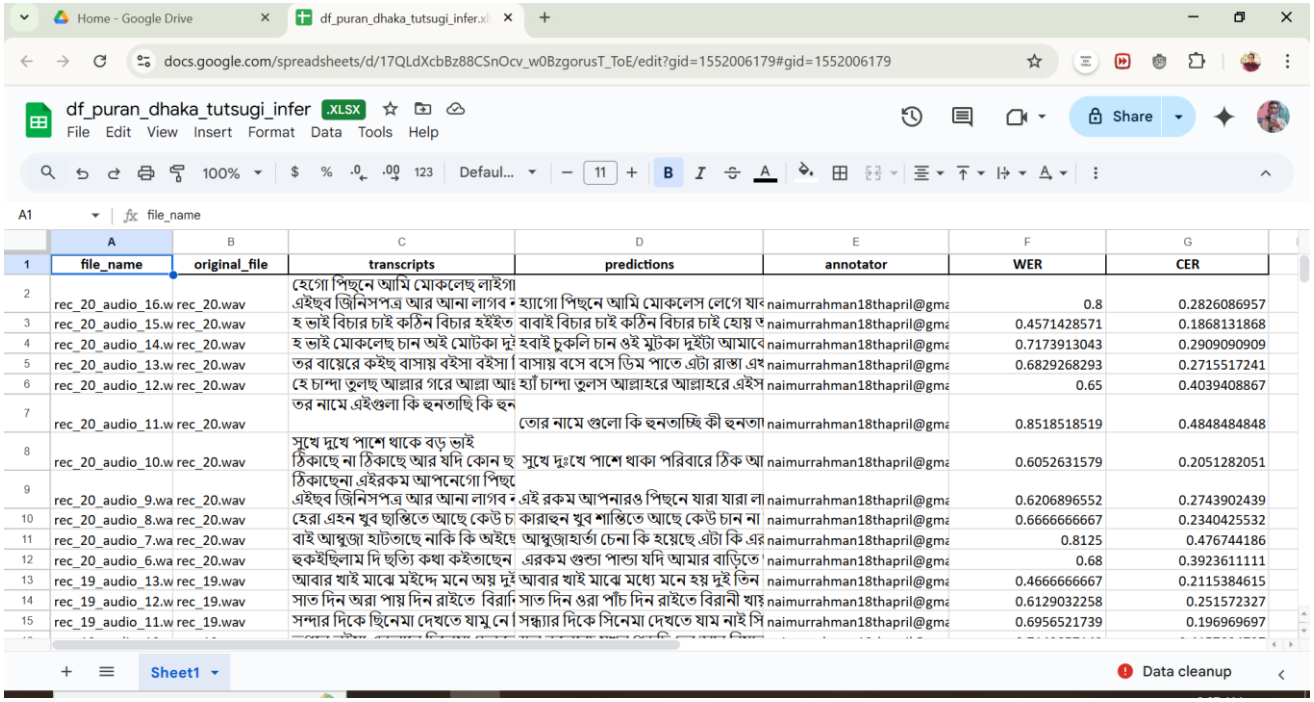
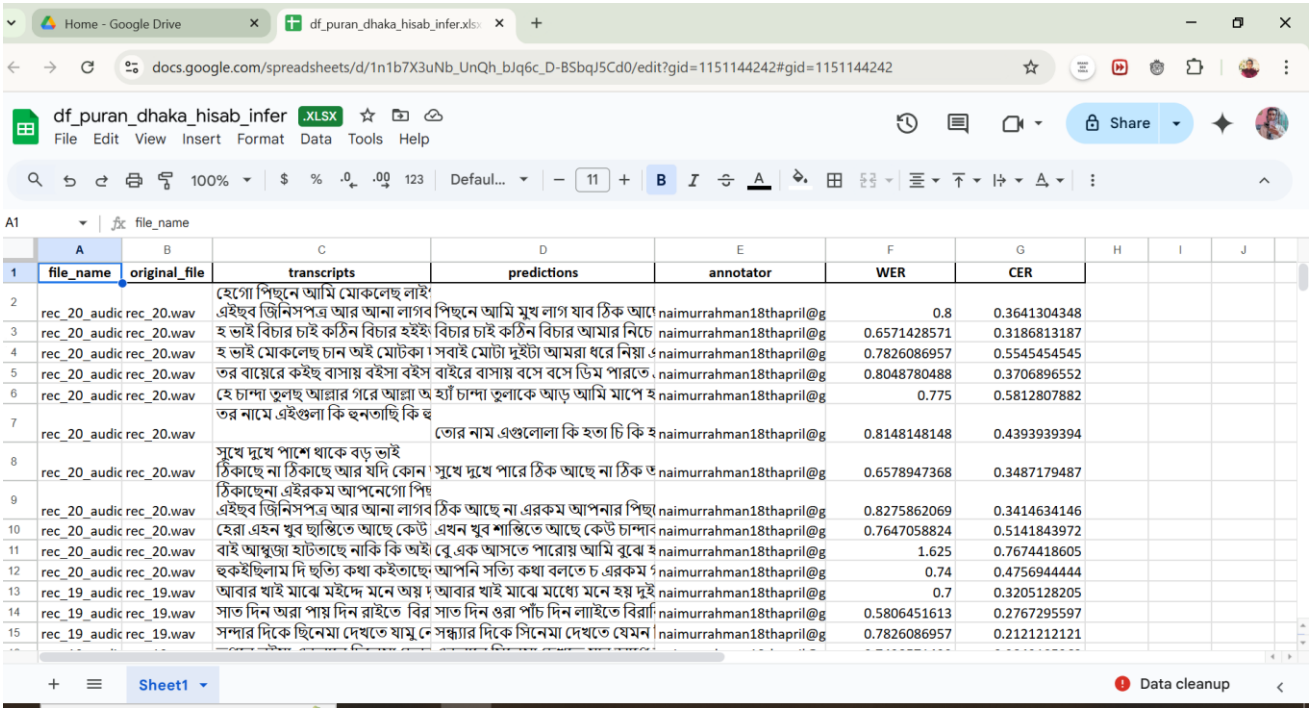


Figure 6.3: Model inferences samples on Puran Dhaka by tugstugi



file_name	original_file	transcripts	predictions	annotator	WER	CER
rec_20_audic	rec_20.wav	হেগো পিছনে আমি মোকলেছ লাই				
rec_20_audic	rec_20.wav	এইছব জিনিসপত্র আর আনা লাগব	পিছনে আমি মুখ লাগ যাব ঠিক আছে	naimurrahman18thapril@g	0.8	0.3641304348
rec_20_audic	rec_20.wav	হ ভাই বিচার চাই কঠিন বিচার হইই	বিচার চাই কঠিন বিচার আমার নিচে	naimurrahman18thapril@g	0.6571428571	0.3186813187
rec_20_audic	rec_20.wav	হ ভাই মোকলেছ চান অই মোটকা	সবাই মোটা দুইটা আমরা ধরে নিয়া	naimurrahman18thapril@g	0.7826086957	0.5545454545
rec_20_audic	rec_20.wav	তর বায়েরে কইছ বাসায় বইসা বইস	বাইরে বাসায় বসে বসে ডিম পারতে	naimurrahman18thapril@g	0.8048780488	0.3706896552
rec_20_audic	rec_20.wav	হে চান্দা তুলছ আল্লার গরে আল্লা অ	হ্যাঁ চান্দা তুলাকে আড় আমি মাপে হ	naimurrahman18thapril@g	0.775	0.5812807882
rec_20_audic	rec_20.wav	তর নামে এইগুলা কি হুনতাছি কি হু	তোর নাম এগুলোলা কি হতা চি কি হ	naimurrahman18thapril@g	0.8148148148	0.4393939394
rec_20_audic	rec_20.wav	সুখে দুখে পাশে থাকে বড় ভাই ঠিকাছে না ঠিকাছে আর যদি কোন	সুখে দুখে পারে ঠিক আছে না ঠিক ত	naimurrahman18thapril@g	0.6578947368	0.3487179487
rec_20_audic	rec_20.wav	ঠিকাছেনা এইরকম আপনেগো পিছ এইছব জিনিসপত্র আর আনা লাগব	ঠিক আছে না এরকম আপনার পিছ	naimurrahman18thapril@g	0.8275862069	0.3414634146
rec_20_audic	rec_20.wav	হেরা এহন খুব ছান্তিতে আছে কেউ	এখন খুব শান্তিতে আছে কেউ চান্দা	naimurrahman18thapril@g	0.7647058824	0.5141843972
rec_20_audic	rec_20.wav	বাই আম্বুজা হাটতাছে নাকি কি অই	বে এক আসতে পারোয় আমি বুঝে হ	naimurrahman18thapril@g	1.625	0.7674418605
rec_20_audic	rec_20.wav	হুকইছিলাম দি ছত্যি কথা কইতাছে	আপনি সত্যি কথা বলতে চ এরকম	naimurrahman18thapril@g	0.74	0.4756944444
rec_19_audic	rec_19.wav	আবার খাই মাঝে মইদ্দে মনে অয়	আবার খাই মাঝে মধ্যে মনে হয় দুই	naimurrahman18thapril@g	0.7	0.3205128205
rec_19_audic	rec_19.wav	সাত দিন অরা পায় দিন রাইতে বির	সাত দিন ওরা পাঁচ দিন লাইতে বিরা	naimurrahman18thapril@g	0.5806451613	0.2767295597
rec_19_audic	rec_19.wav	সন্দার দিকে ছিনেমা দেখতে যামু নে	সন্ধ্যার দিকে সিনেমা দেখতে যেমন	naimurrahman18thapril@g	0.7826086957	0.2121212121

Figure 6.4: Model inferences samples Puran Dhaka by Hishab Conformer

6.3 Benchmarking Performances

The benchmarking performance of different models and our fine-tuned model on our developed regional speech corpus is shown in Table 6.2

Table 6.2: Benchmarking Performance

ASR System	WER	CER
Wav2Vec2 Large	0.921	0.487
Hishab Conformer	0.658	0.349
Tugstugi	0.605	0.205

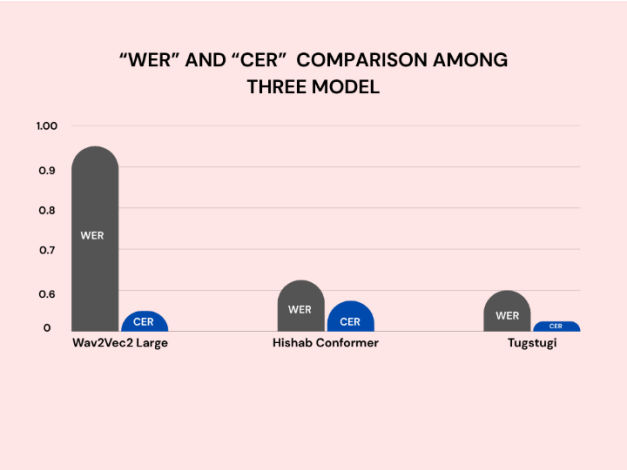


Figure 6.5: Comparison Between Three Model

Wav2Vec2 Large with the given high WER equal 0.921 and CER equal 0.487; it means that the program has serious difficulties in understanding the regional dialect in comparison with other models.

Isolated word accuracy of Hishab Conformer is higher with a WER of 0.658 and CER of 0.349 indicates that Wav2Vec2 Large is more conforming to the dialect of the region.

Among the benchmarked systems, Tugstugi yielded lowest WER of 0.605 and CER of 0.205 making it the best benchmarked system for recognizing the Puran Dhaka dialect.

A sample evaluation has been carried out where Tugstugi finest performance on the regional speech corpus where the lowest error rate was experienced across the two metrics. This has a higher capacity to address the peculiarities of the Puran Dhaka dialect than the other machines compared to the other benchmarked ASR systems.

Comparison of Wav2Vec2 Performance:

In this study, the performance of **Wav2Vec2** was evaluated using **Word Error Rate (WER)** and **Character Error Rate (CER)** for the "Local Speech: A Bengali Regional Speech Recognition Dataset for Benchmarking Under Dialect Variation Puran Dhaka" and compared with results from the "Investigating self-supervised, weakly supervised and fully supervised training approaches for multi-domain automatic speech recognition: a study on Bangladeshi Bangla"

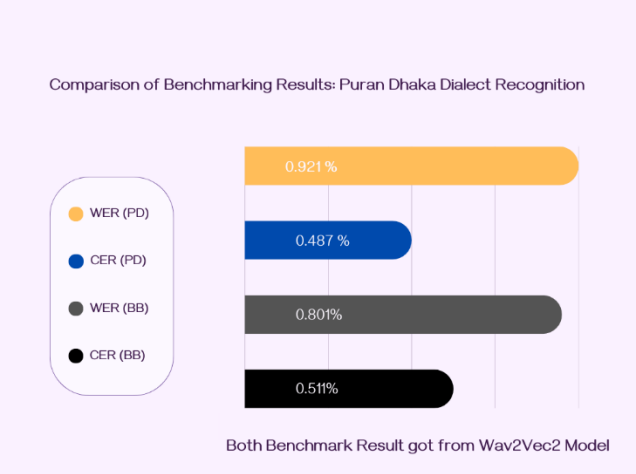


Figure 6.6: Comparison Between Other Research

Chapter 7

Conclusion

The work reported in this paper presents a holistic approach to constructing the first 5 hours of the regional speech corpora designed particularly for the regional varieties of Bengali spoken in Bangladesh. This dataset is the only solution accessible to the public and is focused especially on regional dialects of ASR (Automatic Speech Recognition). We also make an extensive review of the linguistic problems involved in mimicking Bengali speech with regional dialects for which we use created regional speech corpus. However, it is best Suitable for text-to-speech applications, and it offers many opportunities for meta-learning, federated learning and many other applications which are yet to be explored. At present the more refined model, dataset and corpus are still under process of enhancement. The creation of the subsequent versions of the dataset and the model will entail addressing questions such as gender bias and syntactic variability in the corpus.

Also, in an effort to better the understating of regional Bengali type by linguistic researchers, an analysis of regional Bengali language will be conducted. For the purposes of future automated transliteration between standard renderings of Bengali and the local dialects, the informed plans include the canonization of these transcribed data.

Future Work

The foundation for developing Automatic Speech Recognition (ASR) systems that are suited to Bengali dialects specifically, the Puran Dhaka dialect is laid by this research. To expand on the existing work, there are still a few topics that need investigation in the future:

1. **Expansion of the Dataset**

   The dataset for this study includes Puran Dhaka for the first time. To analyse other characteristics of the ASR systems with respect to dialectal differences, future studies could expand the dataset by sampling other regional Bengali dialects.
   Additionally, the community and crowdsourcing can contribute more diverse speech samples for research.

2. **Fine-Tuning and Adaptation**

   Wav2Vec2 Large, Hishab Conformer, Tugstugi et al. could not be properly fine-tuned due to time and GPU resource limits. Thus, future study can promote transfer learning methodologies and optimisation algorithms to improve these models.
   Bengali dialect CER and WER may improve with domain-specific pretraining.

3. **Exploration of Lightweight Models**

   The restriction of resources imply that novel techniques to explore lightweight and effective ASR models for inquiry could be a useful strategy. These models can be trained while possessing remarkable performance even with fewer datasets; they demand less processing power.

4. **Data Augmentation Techniques**

   Exploring new data augmentation approaches may help minimise the restrictions of working with a little amount of data. For example, data diversification with TTS synthesizer or adding noise, variation in speed or pitch to existing test data might enhance the cross generality of the model.

5. **Model Deployment and Real-World Testing**

   By deploying these models in real-world applications such as voice recognition assistants, transcription, or accessibility tools it is likely that the potential of using such tools can be shown. Data collected from such ad hoc implementations can be utilised in further modification of the implementation.

Bibliography

[1] M. Roy, “Some problems of english consonants for a bengali speaker of english,” ELT Journal, vol. 23, no. 3, pp. 268–270, 1969.

[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural ’networks,” vol. 2006, Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[3] F. Alam, M. Habib, D. Sultana, and M. Khan, “Development of annotated bangla speech corpora,” Sep. 2010.

[4] M. M. Rashid, M. A. Hussain, and M. S. Rahman, “Text normalization and diphone preparation for bangla speech synthesis,” Journal of Multimedia, vol. 5, no. 6, p. 551, 2010.

[5] B. Das, S. Mandal, and P. Mitra, Shruti bengali continuous asr speech corpus, 2011. [Online]. Available: https://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html.

[6] S. Mandal, B. Das, P. Mitra, and A. Basu, “Developing bengali speech corpus for phone recognizer using optimum text selection technique,” in 2011 international conference on asian language processing, IEEE, 2011, pp. 268–271.

[7] D. Povey, A. Ghoshal, G. Boulianne, et al., “The kaldi speech recognition toolkit,” IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Jan. 2011.

[8] J. I. Ali, Introduction to phonology. Maola Brothers, Dhaka, 2012.

[9] M. D. Haque, “Bhashabigganer kotha,” 2014.

[10] C. J. Maddison, D. Tarlow, and T. Minka, A* sampling, 2015. arXiv: 1411.0030 [stat.CO]. [Online]. Available: https://arxiv.org/abs/1411.0030.

[11] Bills, Aric, David, Anne, Dubinski, Eyal, et al., Iarpa babel bengali language pack iarpa-babel103b-v0.4b, 2016. DOI: 10 . 35111 / 5jdb - wp44. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016S08.

[12] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Improving non-native mis-pronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling,” in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 6135–6139.

[13] N. D. Londhe, M. K. Ahirwal, and P. Lodha, “Machine learning paradigms for speech recognition of an indian dialect,” in 2016 International Conference on Communication and Signal Processing (ICCSP), 2016, pp. 0780–0786. DOI: 10.1109/ICCSP.2016.7754251.

[14] M. Rashid and S. Chowdhury, “Word sense ambiguity and bangla homographs a linguistic analysis,” The Research Journal of Humanities, vol. 1, pp. 327–36, Jul. 2016.

[15] S. Darjaa, R. Sabo, M. Trnka, M. Rusko, and G. Múcsková, “Automatic recognition of slovak regional dialects,” in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018, pp. 305–308. DOI: 10.1109/DISA.2018.8490639.

[16] C. C. Johny and M. Jansche, “Brahmic schwa-deletion with neural classifiers: Experiments with bengali,” in Proc. The 6th Intl. Workshop on Spoken Lan- guage Technologies for Under-Resourced Languages, 2018, pp. 259–263. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-54.

[17] M. F. Khan, “Construction of large scale isolated word speech corpus in bangla,” Global Journal of Computer Science and Technology, vol. 18, no. G2, pp. 21–26, 2018.

[18] M. Khan and M. Sobhan, “Creation of connected word speech corpus for bangla speech recognition systems,” Asian Journal of Research in Computer Science, pp. 1–6, 2018.

[19] S. Murthy, D. Sitaram, and S. Sitaram, “Effect of TTS Generated Audio on OOV Detection and Word Error Rate in ASR for Low-resource Languages,” in Proc. Interspeech 2018, 2018, pp. 1026–1030. DOI: 10.21437/Interspeech.2018-1555.

[20] TDIL, Bengali speech data – asr, 2018. [Online]. Available: http://tdil-dc.in/index.php?lang=en.

[21] J. Li, V. Lavrukhin, B. Ginsburg, et al., “Jasper: An end-to-end convolutional neural acoustic model,” arXiv preprint arXiv:1904.03288, 2019.

[22] R. Nordquist, Mutual intelligibility, [Online; accessed 23-October-2023], 2019. [Online]. Available: https://www.thoughtco.com/what- is- mutual- intelligibility-1691333.

[23] S. Ahmed, N. Sadeq, S. S. Shubha, M. N. Islam, M. A. Adnan, and M. Z. Is- lam, "Preparation of bangla speech corpus from publicly available audio & text," in Proceedings of The 12th language resources and evaluation conference, 2020, pp. 6586–6592.

[24] R. Ardila, M. Branson, K. Davis, et al., "Common voice: A massively-multilingual speech corpus," English, in Proceedings of the Twelfth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, et al., Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec- 1.520.

[25] N. Choudhary and D. Rao, "The ldc-il speech corpora," in 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2020, pp. 28–32.

[26] A. Ezzine, H. Satori, M. Hamidi, and K. Satori, "Moroccan dialect speech recog- nition system based on cmu sphinxtools," in 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020, pp. 1–5. DOI: 10.1109/ ISCV49265.2020.9204250.

[27] A. Gulati, J. Qin, C.-C. Chiu, et al., "Conformer: Convolution-augmented trans- former for speech recognition," 2020. arXiv: 2005.08100 [eess.AS].

[28] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition," in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 297–301.

[29] S. Kibria, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Acoustic analysis of the speakers' variability for regional accent-affected pronunciation in bangladeshi bangla: A study on sylheti accent," IEEE Access, vol. 8, pp. 35 200–35 221, 2020. DOI: 10.1109/ACCESS.2020.2974799.

[30] H. Liu, J. Liang, V. J. van Heuven, and W. Heeringa, "Vowels and tones as acoustic cues in chinese subregional dialect identification," Speech Communication, vol. 123, pp. 59–69, 2020, ISSN: 0167-6393. DOI: https : / / doi . org / 10 . 1016 / j . specom.2020.06.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639320302417.

[31] I. Nigmatulina, T. Kew, and T. Samardzic, "ASR for non-standardised languages with dialectal variation: The case of Swiss German," in Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), Dec. 2020,
pp. 15–24. [Online]. Available: https://aclanthology.org/2020.vardial-1.2.

[32] S. Shivaprasad S. and M., "Identification of regional dialects of telugu language us- ing text independent speech processing models," in International Journal of Speech Technology volume 23, 2020, pp. 251–258. DOI: https://doi.org/10.1007/s10772- 020 - 09678 - y. [Online]. Available: https:// link .springer.com/ article / 10 . 1007 / s10772-020-09678-y.

[33] M. H. R. Sifat, C. R. Rahman, M. Rafsan, and H. Rahman, "Synthetic error dataset generation mimicking bengali writing pattern," in 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 1363–1366.

[34] R. Smith and T. Rathcke, "Dialectal phonology constrains the phonetics of prominence," Journal of Phonetics, vol. 78, p. 100 934, 2020. DOI: https://doi.org/10.1016/j.wocn.2019.100934. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447018300226.

[35] S. Alam, T. Reasat, A. S. Sushmit, et al., "A large multi-target dataset of com- mon bengali handwritten graphemes," in International Conference on Document Analysis and Recognition, Springer, 2021, pp. 383–398.

[36] J. Lee, K. Kim, and M. Chung, "Korean dialect identification based on intona- tion modeling," in 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2021, pp. 168–173. DOI: 10 . 1109 /O-COCOSDA202152914.2021.9660537.

[37] M. Or Rashid, "How to process bangla linguistic-data through nlp pipeline," Dhaka University Journal of Linguistics, vol. 12, pp. 135–164, Mar. 2021.

[38] S. Sultana, M. S. Rahman, and M. Z. Iqbal, “Recent advancement in speech recognition for bangla: A survey,” Int. J. Adv. Comput. Sci. Appl, vol. 12, no. 3, pp. 546– 552, 2021.

[39] O. Aitoulghazi, A. Jaafari, and A. Mourhir, “Darspeech: An automatic speech recognition system for the moroccan dialect,” in 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), 2022, pp. 1–6. DOI: 10.1109/ISCV54655.2022.9806105.

[40] H. A. Alsayadi, S. Al-Hagree, F. A. Alqasemi, and A. A. Abdelhamid, “Dialectal arabic speech recognition using cnn-lstm based on end-to-end deep learning,” in 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), 2022, pp. 1–8. DOI: 10.1109/eSmarTA56775.2022.9935427.

[41] K. S. Bhogale, A. Raman, T. Javed, et al., “Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages,” arXiv preprint arXiv:2208.12666, 2022.

[42] S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, “Bangladeshi bangla speech corpus for automatic speech recognition research,” Speech Communication, vol. 136, pp. 84–97, 2022.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” arXiv preprint arXiv:2212.04356, 2022.

[44] H. Shahgir, K. S. Sayeed, and T. A. Zaman, “Applying wav2vec2 for speech recognition on bengali common voices dataset,” arXiv preprint arXiv:2209.06581, 2022.

[45] P. Swetha and J. Srilatha, “Applications of speech recognition in the agriculture sector: A review,” ECS Transactions, vol. 107, no. 1, p. 19 377, 2022.

[46] M. R. I. Tomal, T. Kader, A. K. M. Masum, and M. K. A. Chy, “Bangla language dialect classification using machine learning,” in 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), 2022, pp. 1–4. DOI: 10.1109/ICECTE57896.2022.10114552.

[47] M. Wan, J. Ren, M. Ma, Z. Li, R. Cao, and Q. Gao, “Deep neural network based chinese dialect classification,” in 2021 Ninth International Conference on Advanced Cloud and Big Data (CBD), 2022, pp. 207–212. DOI: 10.1109/CBD54617. 2021.00043.

[48] M. Zhai, L. Dong, Y. Qin, and F. Yu, “The research of chain model based on cnn-tdnnf in yulin dialect speech recognition,” in 2022 7th International Conference on Image, Vision and Computing (ICIVC), 2022, pp. 883–888. DOI: 10 . 1109 / ICIVC55077.2022.9886397.

[49] M. Alrehaili, T. Alasmari, and A. Aoalshutayri, “Arabic speech dialect classification using deep learning,” in 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), 2023, pp. 1–5. DOI: 10.1109/ICAISC56366. 2023.10085647.

[50] R. N. Nandi, M. H. Menon, T. A. Muntasir, et al., “Pseudo-labeling for domain-agnostic bangla automatic speech recognition,” arXiv preprint arXiv:2311.03196, 2023.

[51] F. R. Rakib, S. S. Dip, S. Alam, et al., “Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking,” Proc. Interspeech 2023, 2023.

[52] Bengali.AI, Tugstugi$_{b}$*engaliai*−*asr*$_{w}$*hisper*−*medium*(*revisionda*605*cc*), 2024. DOI: 10.57967/hf/2435. [Online]. Available: https://huggingface.co/bengaliAI/tugstugi_bengaliai-asr_whisper-medium.

[53] K. Fatema, F. D. Haider, N. F. Turpa, et al., Ipa transcription of bengali texts, 2024. arXiv: 2403 . 20084 [cs.CL]. [Online]. Available: https:// arxiv . org /abs/

2403.20084.

[54] Aging voice, Accessed: 2023-09-19. [Online]. Available: https:// utswmed . org / conditions-treatments/aging-voice/.

[55] Auditi Das, Wav2vec2-large-xlsr-53-bangla-common*voice*. [Online]. Available: https://huggingface.co/auditi41/wav2vec2-large-xlsr-53-Bangla-Common_Voice.

[56] P. R. Karmaker, "Dialectical and linguistic variations of bangla sounds: Phonemic analysis,"

[57] Labelbox, Accessed: 2024-02-10. [Online]. Available: https://labelbox.com/.

Regional-Speech: A Bengali Speech Recognition Dataset for Benchmarking Models Under Dialect Variation in Puran Dhaka.

ORIGINALITY REPORT

16%	14%	12%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	3%
2	Submitted to islamicuniversity Student Paper	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	arxiv.org Internet Source	1%
5	Bao Thang Ta, Nhat Minh Le, Van Hai Do. "Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
6	aclanthology.org Internet Source	<1%
7	ijeecs.iaescore.com Internet Source	<1%

8	dspace.bracu.ac.bd Internet Source	<1 %
9	pdfs.semanticscholar.org Internet Source	<1 %
10	Submitted to University of New South Wales Student Paper	<1 %
11	helda.helsinki.fi Internet Source	<1 %
12	campus-fryslan.studenttheses.ub.rug.nl Internet Source	<1 %
13	ZhiXing Fan, Jing Li, Aishan Wumaier, Zaokere Kadeer, Abdujelil Abdurahman. "A Multifaceted Approach To Oral Assessment Based On The Conformer Architecture", IEEE Access, 2023 Publication	<1 %
14	Mohammad Muttaqi, Ali Degirmenci, Omer Karal. "US Accent Recognition Using Machine Learning Methods", 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), 2022 Publication	<1 %
15	Ammar Mohammed Ali Alqadasi, Rawad Abdulghafor, Mohd Shahrizal Sunar, Md Sah hj Salam. "Modern Standard Arabic Speech	<1 %

	Corpora: A Systematic Review", IEEE Access, 2023 Publication	
16	researchrepository.universityofgalway.ie Internet Source	<1 %
17	www.ijcaonline.org Internet Source	<1 %
18	hal.science Internet Source	<1 %
19	Harsh Ahlawat, Naveen Aggarwal, Deepti Gupta. "Automatic Speech Recognition: A survey of deep learning techniques and approaches", International Journal of Cognitive Computing in Engineering, 2025 Publication	<1 %
20	ijai.iaescore.com Internet Source	<1 %
21	Nabila Tasnia, Mahidul Islam, Mahi Shahriar Rony, Nishat Tanzim, Khan Md Hasib, Mohammad Shafiul Alam. "An Overview of Bengali Speech Recognition: Methods, Challenges, and Future Direction", 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), 2023 Publication	<1 %
22	knowledgecommons.lakeheadu.ca Internet Source	<1 %

23	Wei Wang, Yanmin Qian. "Universal Cross-Lingual Data Generation for Low Resource ASR", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023 Publication	<1 %
24	huggingface.co Internet Source	<1 %
25	www.cosmosscholars.com Internet Source	<1 %
26	ithesis-ir.su.ac.th Internet Source	<1 %
27	qspace.qu.edu.qa Internet Source	<1 %
28	Sadam Al-Azani, Ridha Almeshari, El-Sayed El-Alfy. "Audio-textual multi-label demographic recognition of Arabic speakers using deep learning", Journal of Intelligent & Fuzzy Systems, 2024 Publication	<1 %
29	d-nb.info Internet Source	<1 %
30	ebin.pub Internet Source	<1 %
31	lib.buet.ac.bd:8080 Internet Source	<1 %

32	Cleyton Aparecido Dim, Nelson Cruz Sampaio Neto, Jefferson Magalhães de Morais. "HornBase: An Audio Dataset of Car Horns in Different Scenarios and Positions", Data in Brief, 2024 Publication	<1 %
33	Xinyuan Song, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Ming Li, Tianyang Wang, Xuanhe Pan, Jiawei Xu. "Transformer: A Survey and Application", Open Science Framework, 2024 Publication	<1 %
34	computerresearch.org Internet Source	<1 %
35	Submitted to Sokoine University of Agriculture Student Paper	<1 %
36	www.jask.or.kr Internet Source	<1 %
37	Noé Tits, Prernna Bhatnagar, Thierry Dutoit. "Text-Independent Phone-to-Audio Alignment Leveraging SSL (TIPAA-SSL) Pre-Trained Model Latent Representation and Knowledge Transfer", Acoustics, 2024 Publication	<1 %
38	onlineresource.ucsy.edu.mm Internet Source	<1 %

39	trepo.tuni.fi Internet Source	<1 %
40	www.manchester.ac.uk Internet Source	<1 %
41	Sadia Sultana, M. Shahidur, M. Zafar. "Recent Advancement in Speech Recognition for Bangla: A Survey", International Journal of Advanced Computer Science and Applications, 2021 Publication	<1 %
42	Wiwit Supriyanti, Sujalwo -, Dimas Aryo Anggoro, Maryam -, Nova Tri Romadloni. "Optimizing Cervical Cancer Diagnosis with Correlation-Based Feature Selection: A Comparative Study of Machine Learning Models", International Journal of Advanced Computer Science and Applications, 2024 Publication	<1 %
43	wrap.warwick.ac.uk Internet Source	<1 %
44	oaktrust.library.tamu.edu Internet Source	<1 %
45	Silin Chen, Tianyang Wang, Xinyuan Song, Bowen Jing, Junjie Yang, Junhao Song, Keyu Chen, Ming Li, Qian Niu, Junyu Liu. "Deep Learning and Machine Learning, Advancing Big Data Analytics and Management:	<1 %

	Generative Models", Open Science Framework, 2024 Publication	
46	Stefano Bini, Vincenzo Carletti, Alessia Saggese, Mario Vento. "Robust speech command recognition in challenging industrial environments", Computer Communications, 2024 Publication	<1 %
47	Shafkat Kibria, Ahnaf Mozib Samin, M. Humayon Kobir, M. Shahidur Rahman, M. Reza Selim, M. Zafar Iqbal. "Bangladeshi Bangla speech corpus for automatic speech recognition research", Speech Communication, 2022 Publication	<1 %
48	lirias.kuleuven.be Internet Source	<1 %
49	www.vut.cz Internet Source	<1 %
50	"Advanced Intelligent Computing in Bioinformatics", Springer Science and Business Media LLC, 2024 Publication	<1 %
51	Submitted to Indian Institute of Technology, Madras Student Paper	<1 %

52	mafiadoc.com Internet Source	<1 %
53	uu.diva-portal.org Internet Source	<1 %
54	graphsearch.epfl.ch Internet Source	<1 %
55	Hae-Sung Jeon, Antje Heinrich. "Perceptual Asymmetry between Pitch Peaks and Valleys", Speech Communication, 2022 Publication	<1 %
56	Submitted to Vrije Universiteit Amsterdam Student Paper	<1 %
57	ds.libol.fpt.edu.vn Internet Source	<1 %
58	9pdf.net Internet Source	<1 %
59	Büchner, Philipp. "Analyse und Spezifizierung der Anforderungen Einer Auf cBioPortal Basierenden Plattform Fuer Molekulare Tumorboards", Friedrich-Alexander-Universitaet Erlangen-Nuernberg (Germany), 2024 Publication	<1 %
60	Sin-wai Chan. "Routledge Encyclopedia of Translation Technology", Routledge, 2023 Publication	<1 %

61	Submitted to University of Wollongong Student Paper	<1 %
62	www.semanticscholar.org Internet Source	<1 %
63	Submitted to Universidad Internacional de la Rioja Student Paper	<1 %
64	ca-cudahy.civicplus.com Internet Source	<1 %
65	vtechworks.lib.vt.edu Internet Source	<1 %
66	www.grafiati.com Internet Source	<1 %
67	www.hindawi.com Internet Source	<1 %
68	www.infinitiresearch.com Internet Source	<1 %
69	"Advances in Smart Medical, IoT & Artificial Intelligence", Springer Science and Business Media LLC, 2024 Publication	<1 %
70	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 Publication	<1 %
71	Submitted to The British College	

	Student Paper	<1 %
72	Xinyuan Song, HSIEH,WEI-CHE, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang. "A Comprehensive Guide to Explainable AI: From Classical Models to LLMs", Open Science Framework, 2024 Publication	<1 %
73	ar5iv.labs.arxiv.org Internet Source	<1 %
74	gadinsider.com Internet Source	<1 %
75	www.frontiersin.org Internet Source	<1 %
76	Jia-Yuan Zhang, Yuning Zhang, Lele Wang, Fei Guo et al. "A single-molecule nanopore sequencing platform", Cold Spring Harbor Laboratory, 2024 Publication	<1 %
77	Liu, Heng. "Algorithms for Scalability and Security in Adversarial Environments.", The University of Arizona, 2021 Publication	<1 %
78	Rami Mohawesh, Shuxiang Xu, Son N. Tran, Robert Ollington, Matthew Springer, Yaser	<1 %

	Jararweh, Sumbal Maqsood. "Fake Reviews Detection: A survey", IEEE Access, 2021 Publication	
79	ia803109.us.archive.org Internet Source	<1 %
80	scyr.kpi.fei.tuke.sk Internet Source	<1 %
81	www.isca-archive.org Internet Source	<1 %
82	1library.org Internet Source	<1 %
83	Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim et al. "Smart reception: An artificial intelligence driven bangla language based receptionist system employing speech, speaker, and face recognition for automating reception services", Engineering Applications of Artificial Intelligence, 2024 Publication	<1 %
84	Konlakorn Wongpatikaseree, Sattaya Singkul, Narit Hnoohom, Sumeth Yuenyong. "Real-Time End-to-End Speech Emotion Recognition with Cross-Domain Adaptation", Big Data and Cognitive Computing, 2022 Publication	<1 %

85	Pingchuan Ma, Stavros Petridis, Maja Pantic. "Visual speech recognition for multiple languages in the wild", Nature Machine Intelligence, 2022 Publication	<1 %
86	Trivedi, Hardi. "Enhancing Cross-Cultural Communication in Low-Resource Language Conversational Agents.", San Jose State University, 2024 Publication	<1 %
87	era.library.ualberta.ca Internet Source	<1 %
88	ltu.diva-portal.org Internet Source	<1 %
89	res.ijsrcseit.com Internet Source	<1 %
90	s-space.snu.ac.kr Internet Source	<1 %
91	unsworks.unsw.edu.au Internet Source	<1 %
92	uokerbala.edu.iq Internet Source	<1 %
93	www.nowpublishers.com Internet Source	<1 %
94	www.researchgate.net Internet Source	

		<1 %
95	www.scirp.org Internet Source	<1 %
96	Dan Jang, Sam Ratnam, Jodi Gilchrist, Manuel Arias, Marek Smieja, Donna Mayne, Max A. Chernesky. "Comparison of Workflow, Maintenance, and Consumables in the GeneXpert Infinity 80 and Panther Instruments While Testing for Chlamydia trachomatis and Neisseria gonorrhoeae", Sexually Transmitted Diseases, 2016 Publication	<1 %
97	"Man-Machine Speech Communication", Springer Science and Business Media LLC, 2024 Publication	<1 %
98	Ahnaf Mozib Samin, M. Humayon Kobir, Md. Mushtaq Shahriyar Rafee, M. Firoz Ahmed et al. "BanSpeech: A Multi-Domain Bangla Speech Recognition Benchmark Toward Robust Performance in Challenging Conditions", IEEE Access, 2024 Publication	<1 %
99	Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, Soujanya Poria. "A	<1 %

review of deep learning techniques for speech processing", Information Fusion, 2023
Publication

100	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1 %
101	Hussain, Shehzeen Samarah. "Robust and Efficient Deep Learning for Multimedia Generation and Recognition", University of California, San Diego, 2023 Publication	<1 %
102	Md. Majedul Islam, Avishek Das, Ibna Kowsar, A K M Shahariar Azad Rabby, Nazmul Hasan, Fuad Rahman. "Towards building a Bangla text recognition solution with a Multi-Headed CNN architecture", 2021 IEEE International Conference on Big Data (Big Data), 2021 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		