

# Examining Lexical and Grammatical Difficulties in Bengali Language using NLP with Machine Learning

by

Sazia Afroz

17301023

M. Samiur Rahman Susmoy

17301155

Farhan Anjum

17301184

Nadiya Nowshin

17301192

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
January 2021

© 2021. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

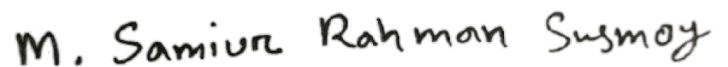
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**



---

Sazia Afroz  
17301023



---

M. Samiur Rahman Susmoy  
17301155



---

Farhan Anjum  
17301184



---

Nadiya Nowshin  
17301192

# Approval

The thesis titled "Examining Lexical and Grammatical Difficulties in Bengali Language using NLP with Machine Learning submitted by

1. Sazia Afroz (17301023)
2. M. Samiur Rahman Susmoy (17301155)
3. Farhan Anjum (17301184)
4. Nadiya Nowshin (17301192)

Of Fall,2020 has been accepted as satisfactory in partial fulfilment of the requirement of the degree of B.Sc. in Computer Science on January 08, 2021.

## Examining Committee:

Supervisor:  
(Member)

---

Mostafijur Rahman Akhond  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Program Coordinator:  
(Member)

---

Dr. Md. Golam Robiul Islam  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

DR. Mahbubul Alam Majumdar  
Chairman  
Department of Computer Science and Engineering  
Brac University

## **Ethics Statement**

The thesis is carried out in complete compliance with research ethics, policies, regulations and codes set by BRAC University. We have used various information from different sources in order to pursue the research. To collect data, we read articles, journals from different websites etc. The sources we have used here are interpreted in our own terms and are properly mentioned as reference. We appreciate and give credit to every source that helped us to continue our work. Lastly, we declare that four authors of this paper hold liability if any violation of BRAC university standard is found.

## Abstract

Speech is the most relevant form of communication between individuals whereas computer systems use basic form of text or images. Speech synthesis deals with the artificial production of speech, and a Speech-to-Text (STT) system in this aspect converts natural dialogue into a corresponding written language. There have been sufficient successes today in the area of speech and natural language processing that suggests that these technologies will continue to be a significant area of research and development in creating intelligent systems. In this paperwork, we are planning to provide an overview of the STT synthesis technology along with details of the phases involved. For conversion of speech the proposed model is pocketsphinx which was trained to generate an acoustic language model. Additionally, our system will enable us to detect the grammatical and contextual errors of the texts. The thrust has been given to explore the usefulness of this technique in designing an STT the above system for Bangla language. This paperwork will also focus on some of the open research issues where work in this area may further be done.

**Keywords:** STT; NLP; HMM; PocketSphinx; Speech recognizer; Grammatical analyzer; Contextual analyzer

## **Dedication**

We would like to dedicate our thesis to our parents who gave support by all means to come this far. Then to our respected supervisor Mostafijur Rahman Akhond sir and co-supervisor Annajiat Alim Rasel sir. We could not have conducted our research without their instructions and guidelines. Lastly, with condolence, to all the people who had and still are suffering loss from data stealing and identity thieving issues.

## **Acknowledgement**

The name of almighty who gave us the strength, we are grateful to our family members. As well as to our supervisor and co-supervisor, the people who kept guiding us throughout the whole time of this thesis. We also appreciate our fellow team members who hold strong and united till the end to successfully complete the work. Also, our friends who helped us to remain less stressful during this pandemic. Lastly, to our university and it's authority, where we got the platform to approach one step forward to achieve our goals.

# Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iii
Abstract	iv
Dedication	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Problem Statement . . . . .	2
1.3 Research Objective . . . . .	3
<b>2 Literature Review</b>	<b>4</b>
<b>3 Proposed System</b>	<b>8</b>
3.1 Speech-to-text . . . . .	9
3.2 Grammatical analyzer . . . . .	11
<b>4 Implementation and Result Analysis</b>	<b>16</b>
4.1 Speech Recognizer . . . . .	16
4.1.1 Data Preparation . . . . .	16
4.1.2 Experimental Result and Analysis . . . . .	17
4.2 Grammatical analyzer . . . . .	19
4.2.1 Data Preparation and Implementation . . . . .	19
<b>5 Challenges</b>	<b>23</b>
<b>6 Conclusion</b>	<b>24</b>
6.1 Conclusion . . . . .	24
6.2 Future Work . . . . .	24



<b>Bibliography</b>	<b>25</b>
Bibliography . . . . .	25

# List of Figures

3.1	The HMM topology used in SPHINX . . . . .	8
3.2	System Architecture . . . . .	9
3.3	Language model . . . . .	10
3.4	Dictionary . . . . .	11
3.5	Speech recognizer block diagram . . . . .	11
3.6	Tokenization . . . . .	12
3.7	Parts of Speech tags . . . . .	13
3.8	Word-tag tuples . . . . .	13
3.9	Viterbi Algorithm example . . . . .	14
3.10	Tag sequence . . . . .	14
3.11	Application Workflow . . . . .	15
4.1	File id . . . . .	17
4.2	Transcription file . . . . .	17
4.3	Table with number of users to SER and WER . . . . .	18
4.4	Tag sequence . . . . .	20
4.5	Word tag vocabulary . . . . .	20
4.6	Probability Equation . . . . .	21
4.7	Example of count matrix . . . . .	21
4.8	Example of sum matrix . . . . .	21
4.9	Example of probability matrix . . . . .	22

# Chapter 1

## Introduction

### 1.1 Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence that is mainly focused on communication between computers and human languages. It is rather difficult to teach computers the linguistics of human language, but in recent years, we have made several breakthroughs in this sector. The primary purpose of NLP is to be able to read, decipher, and categorize various human languages.

Most people square measure fluent in speaking language; however, their writing talent is appalling because of their lack of grammatical knowledge and oversight at the time of writing. With the advancement of procedure algorithms, the expectation of individuals is also increasing day by day. Rather than depending only on mechanical assistance, we are now seeking intellectual support as well. Millions of people deal with texts throughout the world without having proper knowledge about the language, and most of them are not necessarily native speakers.

Millions of individuals trot out texts throughout the planet while not having correct grammar concerning the language, and many of them are not essentially native speakers. Most of them use writing system correction tools while writing documents. These tools offer a primary step towards writing correct text by saving massive human intervention. The second step of high-quality text generation is descriptive linguistics checking. Descriptive linguistics checking is essential for many reasons. It improves the standard of the text, saves time, and supports the training of the language. This kind of tool not solely helps native speakers but also helps the second language learners to speak in an alternative language. STT technology is one of the most important functions for converting human language into a machine-readable format. It is also known as "speech recognition", "automatic speech recognition," or simply "computer speech recognition". Speech is a continuous stream of audio where dynamically changed states are in between stable states. Speech data also contain sounds and phones which are non-linguistic sounds. These less similar classes of sounds are called fillers for example breath, um, uh, cough. They are separate chunks of audio between pauses. They do not necessarily match sentences, which are more semantic concepts [1,2,3].

Recently, conversion between speech and text speech-oriented human-computer in-

teractions is highly attracting researchers. The transfer of speech into text in real-time requires special techniques because it must be almost simultaneous, and almost 100 percent correct to be understandable. Continuous speech recognizer is complex to build as they utilize special techniques and unique sounds to determine utterance boundaries. Moreover, the environmental variability, speaker style, gender, accent, age, speed of speech also makes the system more complex. Recognition rate can vary among paragraphs as word repetitions may be different may be different if the recognition is done word-wise. Also, there are some sentences composed of idioms and phrases, and mixed sentences are difficult to handle for the system.

The importance of STT is recognized, and various toolkits are produced to make speech recognition much more comfortable. The most popular ones include CMUSphinx[1,2,3], Microsoft Speech Application Programming Interface (SAPI) [5], and Kyoto Julius [7]. This paperwork focuses on the conversion of speech to its corresponding text in Bangla language using the knowledge of NLP. Additionally, the generated text would be analyzed for grammatical errors as well as a contextual error with the help of grammatical analysis and contextual analysis. Our main objective is to propose a framework to detect grammatical and contextual errors in Bangla sentences which are mostly generated from the speech to text converting algorithms. The intended model is capable to produce texts the speech automatically, along with contextual and grammatical error detection functions. The main application of this system is to use on platforms where continuous speech has to be recorded in writing, such as in news interview sessions where everything is handwritten, which has been proven overwhelming for the reporter to note down. In places like these, the written data must be accurate as well as valid and should make sense both grammatically and conventionally.

## 1.2 Problem Statement

For English there has an electronic syntax checker yet in the instance of Bangla this can't set up. Our objective is to structure a Bangla Grammar error detection system that does not just discover the syntactic mistakes yet besides locating the semantic blunders of a sentence. Word expectation strategy predicts word by investigating past word stream for auto finishing a sentence with more precision by sparing the greatest keystroke of any client or understudy. Right now, we propose a strategy of mistake recognition that can pinpoint the blunder position in a major dominant part of cases and this manner decreases the number of right options in contrast to a huge degree.

Using viterbi after identifying the constituents, the productions for viterbi are developed for Bangla sentence structures. As Bangla grammar has different forms, the same production term can be used only by reorganizing the words in the grammar. For example, the following three forms can be applied by reorganizing the production terms.

1. আমার নাম ফারহান ।
2. আমার ফারহান নাম ।
3. ফারহান আমার নাম ॥

The parser created from this kind of linguistic use is not productive because it requires backtracking. To evacuate the uncertainty from the linguistic use the thought of cleared out calculating can be used and reconstruct linguistic use preparations. Cleared out figuring could be a linguistic use change valuable for creating a linguistic use appropriate for prescient parsing. Besides, word expectation implies speculating the following word in a sentence. Such as,

আম।

আমরা

আমড়া

আশ্মা

But there is no satisfactory analysis of the Bangla language to predict words in a sentence. So, some N-gram language models, back off, and deleted interpolation techniques can be applied to predict Bangla words in a sentence. Large data set of text words in Bangla can also be used.

### 1.3 Research Objective

The study of grammatical and contextual errors of the texts of the Bangla language enables the communities to interpret and perform specific tasks if the orders, sentences, or symbols are meaningful. Our main objective is to describe a methodology to detect lexical and grammatical errors from Bangla sentences.

- Our target is to compare past systems with our proposed systems
- The main application of this system is to use on platforms where continuous speech has to be recorded in writing, such as in court sessions where everything is handwritten but it has proven too difficult for human hands.
- Our main motive is to digitalize this sort of platform where speech to text is automatically done along with lexical and grammatical error detection function.
- After all these our objective is to observe how people speak and react under different scenarios, such as meeting rooms vs. open area, single speaker vs. multiple speakers, and so on.

Finally, our objective is to simulate the worst scenario in the virtual environment so that our system can work fine in the real world.

# Chapter 2

## Literature Review

The key prerequisite for this documentation is to go through numerous toolkits that are required for Speech-to-text conversion. Kyoto Julius is one of the speech recognition engines we have come across. Julius is an open-source continuous speech recognition engine with broad vocabulary. It's supported word N-gram and context-dependent HMM. On different devices, ranging from microcomputers to cloud servers, Julius can perform real-time decoding. The algorithm is based on two searches for tree trellis passes. For instance, the tree-organized lexicon, N-gram factoring, Gaussian choice, etc. It is also independent of model structures, and a large range of HMM structures are provided with any number of mixtures, states, or phone sets as mutual state triphones and tied-mixture models. In addition, on one line, Julius can run multi-instance dictation recognition, grammar-based recognition, or isolated word recognition. In addition, common formats such as The Hidden Markov Model Toolkit (HTK), The SRI Language Modeling Toolkit (SRILM), Deep Neural Network (DNN) based on real-time decoding are used for models to deal with other speech/language modeling toolkits. Linux is its primary platform; however, in windows, Mac, Androids, it is also supported [1].

Furthermore, the Java-based toolkit we tested was CMUSphinx-4. CMUSphinx-4 is a modular and pluggable architecture that integrates existing device design trends with ample versatility to serve emerging areas of interest in science. The framework is modular in that it contains separable components dedicated to particular tasks, and it is easy to substitute its pluggable modules at runtime. Sphinx-4 also contains a distribution of modules that incorporate state-of-the-art speech recognition techniques in order to exercise the platform and to produce researchers with a working system[2].

A BRAC University research paper by Shammur Absar Chowdhury and Prof. Mumit Khan on Continuous Speech Recognition (CSR) in Bangla says right now that the best way to use CMUSphinx[1] tools to create an area-based ceaseless discourse recognition framework is to prepare, prepare and incorporate content into the discourse corpus with the framework[3].

At Shahjalal University of Science and Technology, a further analysis of this work was performed. In this analysis, they believe that the performance of the system using CMU Sphinx4[1,3] would be enough to use the system in practical life if they

can increase the training data. Their accuracy rate is 85%. Using a smartphone where their accuracy rate is 75%, they have tested the device. The proposed method was tested with only 115 vocabulary size training data which is very small amount and is not enough[4].

They are developing a speaker-independent recognition device for Continuous Speech, another study under the supervision of Hossain Arif. They used PocketSphinx to create a speech recognition system[1]. With the aid of PocketSphinx, an acoustic model has been educated, a language model has been developed, and a dictionary is needed to convert speech to text. The second element is the Dictionary, which includes the words and the phoneme of their pronunciation. It includes a mapping of vocabulary words to phoneme sequences for the system.

Acoustic model, which defines the language's sounds. It offers a diagram between a speech unit and an HMM. After creating a language model, dictionary, fields, and transcription, they wrote a python command.

The input voice will initially be dissected to form individual words in silence. Then the detected phone sets will cross-refer to the acoustic model, the language model, and the dictionary for each sample and print the matched word. This method repeats iteratively until it has been translated to text for the full length of the given input waveform. The first step in measuring the error in their system was to record unknown and known speakers to the system with Bengali sentences. And, in a separate folder, they placed all the recordings. After that, a test file was generated where the directory and filename of each test recording were specified. Then they also created a test transcription file where they listed the filenames of all test recordings against the phrases documented in the recordings. They started decoding the device with CMUSphinx after it was all set. The acoustic model they learned and the language model they configured used the decoding process. When the recognition job was complete, the system generated the output from each of our test recordings[5].

Microsoft Organization created SAPI for speech-related works in its Windows working frameworks. A consider in Khulna College explored STT change utilizing SAPI for Bangla dialect. Here English dialect is utilized as a middleware to oversee SAPI for Bangla STT transformation. To recognize Bangla's elocution, for each Bangla word a combination of English character set is produced into an XML linguistic use record. The XML language structure record is at that point stacked into the SAPI, which at that point returns an English character set on the off chance that the talked word matches. Be that as it may, the issue with SAPI is that it works consecutively which makes it moderate. Within the current methodology, Bangla discourse is recognized on a word by word basis. An person need to conversation with a authentic break in each word so that framework composes the word on the off chance that a coordinate happens[6].

A study by K.M. Azharul Hasan, Al-Mahmud, and Amit Saha, at Khulna College of Designing and innovation, is around utilizing CFG for Bangla dialect and subsequently building a Bangla parser subordinate on the sentence structure. The proposed parser may be a prescient parser and the parse table has been developed for

recognizing grammatical botches of Bangla sentences. All sorts of conventional and non-traditional Bangla sentences can be identified by the displayed plot successfully. Greatest basic, complex, and compound sentences can be secured with this auxiliary representation. But sentences composed of expressions and expressions, too blended sentences are past the scope of this[7]. On a diverse consider at the same college in December 2014, they proposed a strategy for semantic rightness of basic Bangla sentences. The calculation can effortlessly be amplified for complex and compound sentences. The semantic of a Bangla sentence depends on the verb(s) utilized within the sentence. They have categorized the words, so that, each of the words within the same category has the same semantic relationship with other substances. Sentences are broken into a subject, protest, and verb parts for the reason of categorization.

Utilizing the subject-verb relationship, they have made a Approval Table (VT) to check the semantic acknowledgment. The passages of VT are a Boolean relationship Genuine (T) or Untrue (F). It is imperative and fundamental to total the approval table and question verb connection table for all the verbs and things of the Bangla dialect. The execution of the proposed procedure significantly depends on this. In spite of the fact that Bangla could be a complex dialect with a complex structure, in this manner, they have confronted numerous challenges in this technique[8]. Another paper published in 2015, is based on word forecast on Bangla sentence by utilizing stochastic, i.e. N-gram dialect show like unigram, bigram, trigram, erased Introduction, and back off models for auto-completing a sentence by anticipating an precise word in a sentence which spares time and keystrokes of writing conjointly diminishes incorrect spelling. They utilize a huge information corpus of Bangla dialect of different word types to anticipate the right word with the exactness of the most extreme sum as conceivable.

They utilize a factual forecast strategy such as an N-gram procedure for case unigram, bigram, trigram, bakeoff proliferation, erased introduction. They too utilize a expansive information set of content words in Bangla which is collected from distinctive daily papers. N-gram dialect models give a characteristic approach to the development of sentence completion frameworks. To degree the probabilities by a factual demonstrate like N-gram information is part into a preparing set and test set. They have developed N-gram models of word forecast by tallying frequencies of words in an awfully expansive corpus, i.e. database, and decide probabilities utilizing N-gram[9].

There's another promising work from the Indian Established of Innovation. Here they select k best redress sentences from the trellis and after that they utilize the adjusted BLEU (Bilingual Assessment Understudy) Score and WER (Word Mistake Rate) for selecting a correct sentence from multiple correct sentences to guarantee the right sentence isn't as well far from the ungrammatical sentence. Clearly, utilizing the tall BLEU score and low WER they select the right sentence from a set of rectify sentences to guarantee that the right sentence isn't as well distant from the ungrammatical sentence. Other than, their dialect show is prepared on around 4 lakh 80 thousand collected well-formed Bangla sentences and tried on a subset of sentences containing auto-generated ill-formed and collected well-formed sentences. They too utilize a few determination rules as in case a anticipated redress sentence



has an antonym of the writer's postposition, and after that they must not select the sentence in arrange to dodge diverse implications[10].

A research by Arora, Kaur and Gupta from Stanford College utilized the neural arrange acts as a multivariate classifier, and yields probabilities for each target word given the source sentence. NNGLM calculates the likelihood of person plausibility of words by preparing a thickness estimation show. To speed up the computationally costly preparing of neural network's demonstrate, they utilized Python's bundle Pytorch for adaptability and speed. The preparing set contains one million words and is totally clarified with mistake labels and redresses. In arrange to plan a reasonable engineering for NNGLM that can accomplish great exactness for the errand of GEC, they start by preparing a single-layer neural arrange. They begin preparing the show with number of covered up layer neurons 1000, bunch measure 15 and with learning rates 0.02, 0.1 and 0.5 [13].

Kundu from Indian Organized of Innovation performed a think about where they select k best redress sentences from the trellis and after that they utilize adjusted BLEU (Bilingual Assessment Understudy) Score and WER(Word Blunder Rate) for selecting redress sentence from numerous adjust sentences to guarantee rectify sentence isn't as well distant from the ungrammatical sentence. Clearly, utilizing the high BLEU score and moo WER they select the right sentence from a set of redress sentences to guarantee that redress sentence isn't as well distant from the ungrammatical sentence. Other than, their dialect show is prepared on roughly 4 lakh 80 thousand collected well-formed Bangla sentences and tried on a subset of sentences containing auto created ill-formed and collected well-formed sentences. They moreover utilize some selection rules like in case a anticipated redress sentence has an antonym of the writer's postposition, and after that they must not select the sentence in arrange to dodge diverse meaning[14].

A study from Brac College by Humayun Kabir and Ruhan Ahmed,Abdullah Umar Nasib where they made two calculations to rectify word spellings that show up the same but are immensely distinctive in arrange to get the foremost one of a kind frame of each word as well as their Banglish comparable in arrange to create the phone sets. The special UTF word finding calculation remedies the spelling botches by changing over them to "Banglish" to unify the spelling and after that supplanting them with the proper form. This can be performed iteratively for each variety of each word to as it were have a single interesting spelling for each at the end. Moreover an open source outline work called CMU Sphinx 4 was utilized to produce Bengali UNICODE text style. Also digital sound workstation called Dauntlessness was utilized to control the recorded information. The execution of the proposed show was tried employing a dataset where both male and female voices were recorded. The proposed demonstrate appeared around 75% precision for the tried dataset [15].

# Chapter 3

## Proposed System

Our Proposed system is composed of three connected layers. The first layer is the data acquisition layer. The user speech can be captured in this portion using various sensors. The second layer is the data storage layer which stores voice data as well as trained data sets. We plan to use big data models like MapR, IBM storwize, etc in the future.

The third layer is our main application layer. We will use the CMUSphinx framework [1], which is developed based on the Hidden Markov Model (HMM). CMUSphinx is viewed as a tremendously standard strategy for acknowledgment as it is a productive calculation. This plan inside the structure is diverse contrasted with past executions as it builds a diagram that empowers parallelism of unraveling at various levels. This permits synchronous element acknowledgment.

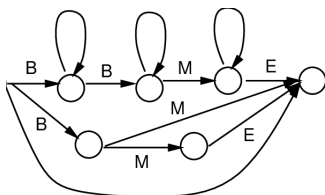


Figure 3.1: The HMM topology used in SPHINX

Word expectation strategy predicts word by investigating past word stream for auto finishing a sentence with more precision by sparing the greatest keystroke of any client or understudy. Right now, we propose a strategy of mistake recognition that can pinpoint the blunder position in a major dominant part of cases and this manner decreases the number of right options in contrast to a huge degree.

At first, we need to collect recordings of different speakers for unique words. Then according to those unique words, we will add a phone against those specific Bengali words in the dictionary file. Furthermore, we will also store these unique phones along with their corresponding dictionary words in the data set storage. To make an understanding of the system, we need to create a bridge so that the system could recognize the words for individual speakers and can match it in the testing phase.

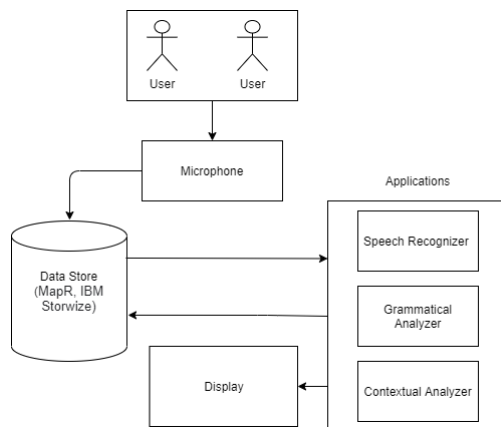


Figure 3.2: System Architecture

The above model describes the system function of our proposed model. The functionalities can be best explained in terms of users. When the user, for example, a reporter or an interviewee, speaks, the microphone captures the articulation, which is then stored in the primary data store. Afterward, the audio recording is passed through our application. The application contains the speech recognition system as well as the algorithm for grammatical analysis and contextual analysis. The speech recognition system parses the audio recording and converts it into text. These converted texts are then checked for grammatical errors and contextual errors.

Our system can be divided into two parts. These divisions are described below:

### 3.1 Speech-to-text

The primary objective of the system is to convert human language into text format provided it is a continuous audio stream. Most of the work in language database model is done in English but less are done in Bangla which makes it difficult to build an acoustic model for Bangla. For the proposed model, we trained an acoustic model with our recorded data where this acoustic model works in a series of steps using the framework named CMUSphinx[1,2,3].

For the proposed system latest version of the following libraries are used.

- Pocketsphinx: A lightweight recognizer library written in C
- Sphinxbase: Supports library for Pocketsphinx
- Sphinxtrain: An acoustic model training tool

The CMUSphinx framework relies on Hidden Markov Model (HMM). It has three key models named acoustic model (AM), language model (LM) and phonetic dictionary file which are used for detecting any language. The decoder in this system has a software, that performs the conversation from speech to text and in our project, we are using PocketSphinx for that.

Language Model is the first an important component of speech recognition system because it helps decoder to recognize sequence of possible words. It describes

the probability of occurrence of a given word based on n-1 previous words. For the proposed model to generate the probability, we build a statistical language model which represents the structural constraints available in the language and specifies the valid words in the language also their arrival sequence in the speech data. The following figure shows a sample of our language model with its statistical values.

Probability	Words
-0.3006	অনর্থ
-0.2986	অনেক
-0.3006	অঙ্ককার
-0.3006	অপ্যায়ের
-0.3006	অপর
-0.2612	অপেক্ষা
-0.3006	অর্থ
-0.3006	অহংকার
-0.3001	আকাশ
-0.3006	আকাশটা

Figure 3.3: Language model

The second component is the Dictionary file containing words and their pronunciation phoneme which provides the system with a mapping of vocabulary words to sequence of phonemes. In this file we manually added all the phonemes for each words and each line contains a unique Bengali word followed by a tab in the actual text file and then its respective phone sets sequence separated by space. If the input file does not find a match, it will search in the language model to discover the closest match.

The third component is the Acoustic Model for describing the sound of the language which provide a mapping between a unit of speech and an HMM. As there is no existing acoustic model present in Bengali that's why we had to build and train an acoustic model.

Words	Phones
অনর্থ	অনর্থ
অনেক	অনেক
অঙ্ককার	অঙ্ককার
অন্যায়ের	অন্যায়ের
অপর	অপর
অপেক্ষা	অপেক্ষা
অর্থ	অর্থ
অহংকার	অহংকার
আকাশ	আকাশ
আকাশটা	আকাশটা

Figure 3.4: Dictionary

Accordingly a detection model is working starting from voice input to text output shown. To form individual word at first the input voice will be dissected at silence. Therefore detected phone sets for each sample will cross reference from acoustic model, the language model and dictionary to print the matched word.

There's a special phone called SIL present no matter which language is required to be detected as it represents the silenced parts of an audio recording. This is also a fundamental file to get right as it affects the final accuracy of detection. It is not clear what the individual phones of a word may sound like in broken form and sometimes may not possible to be automated by the dictionary tool as it can lead to very bad WER(Word Error Rate).

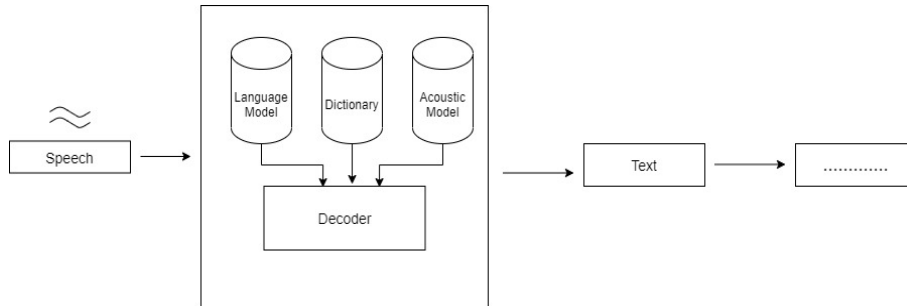


Figure 3.5: Speech recognizer block diagram

## 3.2 Grammatical analyzer

Developing or using the analysis to check the sentence structure of the given audio stream. This part ensures that the sentence is grammatically correct. The sentence must have a subject-verb agreement along with other grammatical laws that are within the Bangla language fundamentals. The primary idea is to check the probability of a given sequence of words to identify whether the sequence of individual

words can form a sentence.

Initially, the speech to text generated sentence is tokenized. The sequence of strings is broken up into pieces that is words which are called tokens. Tokenization plays a large part in the process of lexical analysis. Tokenization relies mostly on simple heuristics in order to separate tokens. The tokens are separated by white spaces. White spaces are not included in the tokens. All the characters in the contiguous strings are part of the tokens.

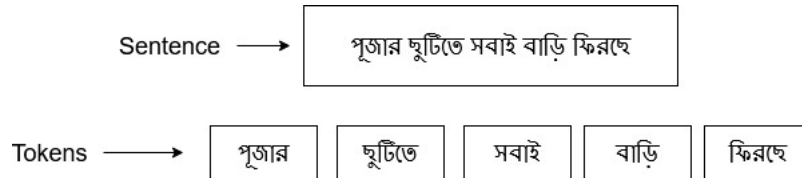


Figure 3.6: Tokenization

After the tokenization process, the tokens are then tagged into their parts of speech corresponding to the original string sentence. Part-of-speech tagging is a process of converting the tokens into a list of tuples where each tuple is in the form (word, tag). The word is the token and tag is its part of speech which signifies whether the word is a noun, verb, adverb etc. Additionally a tag sequence is generated for the corresponding sentence.

Parts Of Speech	Tags
Adjective	J
Noun	N
Conjunction	C
Verb	V
Pronoun	P
Adverb	D
Preposition	PP

Figure 3.7: Parts of Speech tags

In the figure, it is shown the parts of speech which has been used to tag sentences.

(পূজার, "N") (ছুটিতে, "N") (সবাই, "P") (বাড়ি, "N") (ফিরছে, "V")

Figure 3.8: Word-tag tuples

The arrangements of word-tag is then used to populate the progress matrices, which holds the probabilities of moving between various states in the Markov model. These probabilities are utilized to produce change networks. Next, the arrangement can be processed utilizing the Viterbi algorithm to discover the succession of concealed states or grammatical forms labels that have the most elevated likelihood for this grouping. The Viterbi algorithm really processes a few ways simultaneously to locate the most probable grouping of concealed states. It utilizes the framework portrayal of the Hidden Markov model. The calculation can be part into three primary advances: the initialization step, the forward pass, and the backward pass.

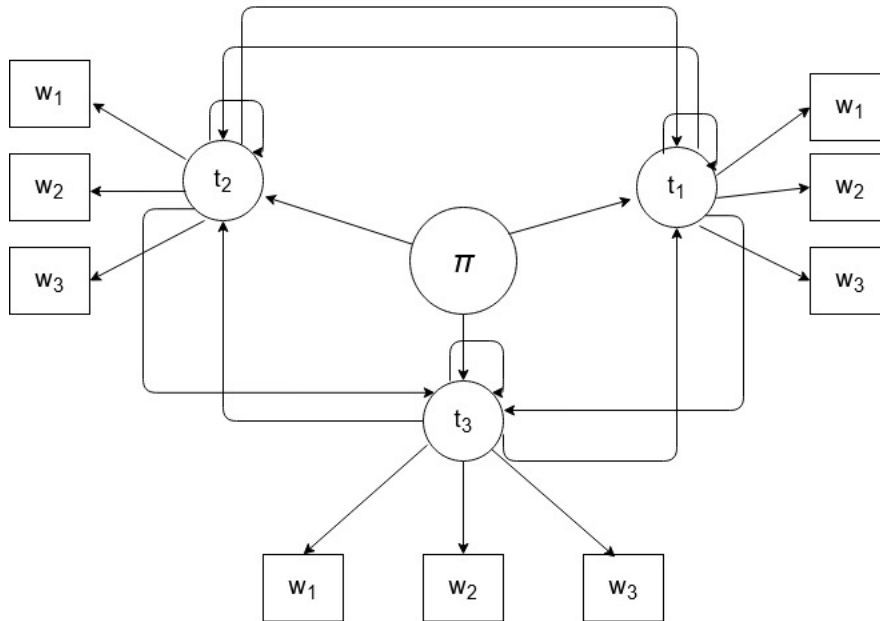


Figure 3.9: Viterbi Algorithm example

***NNPNV***

Figure 3.10: Tag sequence

With the tag sequence, we can now predict the sequence of parts of speech tags with the highest probability. The input sentence has the string of sentence along with its parts of speech tags. The model can now predict whether the string of sentence has the highest probability of tag sequence.

The flowchart of the Application workflow is as follows:



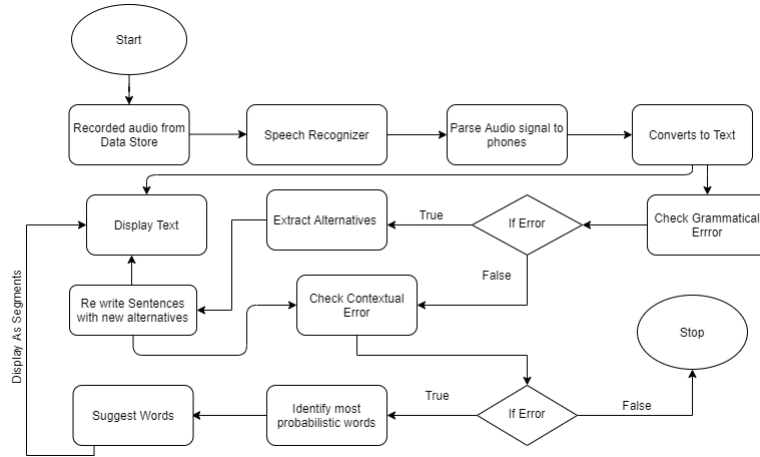


Figure 3.11: Application Workflow

The recorded text is passed through the speech recognizer which parses and converts the audio file into corresponding Bangla text. This converted text is then passed through a grammatical analyzer to check for error. If an error is found, correction is suggested then passed through the contextual analyzer. The contextual analyzer checks if the context is meaningful and based on the error suggestions are displayed.

**Examples of predicted texts:**

Input: তুমি আকাশে উড়ে

Predicted Text: পাখি আকাশে উড়ে

Input: আমি কুলে যায়

Predicted Text: আমি কুলে যাই

সে কুলে যায়

# Chapter 4

## Implementation and Result Analysis

### 4.1 Speech Recognizer

#### 4.1.1 Data Preparation

For building a speech recognizer we needed lots of data which was not easy to collect in this short period of time. So for this we recorded 80 training set sentence and 20 test set sentence, containing 252 unique words. A total number of 10 speakers were asked to record each sentences in different environment for creating more diversified and accurate results.

There is a phone set file which contains every unique phone used for every word in the dictionary. These phones are the essential building block of syllables of words in a recording. Phone division in Bangla language is rather a complex process. That is why to simplify the algorithm we used the entire word as a phone for its corresponding original word. However, this limited us to using limited number of words as total number of phones restricted to 255 phones. Moreover there is a special phone called SIL used for detecting silenced part of the sentence. SIL plays a vital role in generating good word error rate (WER) as it accounts for broken form of words and failed utterances.

The file id keeps track of all the audio files used for training and testing in specific folders. The transcription file uses this to map its sentence with the provided audio wave file. Audio wave files are formatted into distinct specification with a bit resolution of 16 bit, sampling rate 16 KHz and audio channel mono. Another importance of file id is also used to keep track and convert the files into MFC format (Mel-Frequency Cepstrum) representing the short term power spectrum of a sound required for training efficiency during the training and testing process. For creating this file we wrote a script in python. This file shows the list of the audio files where every speaker recordings are sorted together and named in numbers from 1001. The MSB number signifies the speaker id that is “1” and the rest numbers are the particular sentence number “001”

1	Male1/1001
2	Male1/1002
3	Male1/1003
4	Male1/1004
5	Male1/1005
6	Male1/1006
7	Male1/1007
8	Male1/1008
9	Male1/1009
10	Male1/1010

Figure 4.1: File id

The transcription files for testing and training represents sentences to its corresponding audio file ids which is used to train and test the HMM model. The transcription file contains a starting tag, string of sentences followed by an ending tag.

1	<s>	পাখিটি সুন্দর	</s>	(Male1/1001)
2	<s>	গাছে ফুল ফোটে	</s>	(Male1/1002)
3	<s>	রাখাল বাশি বাজায়	</s>	(Male1/1003)
4	<s>	মাঝি নৌকা চালায়	</s>	(Male1/1004)
5	<s>	আকাশটা নীল	</s>	(Male1/1005)
6	<s>	চল ঘুরে আসি	</s>	(Male1/1006)
7	<s>	বাতাস বইছে	</s>	(Male1/1007)
8	<s>	আমার শীতকাল ভালো লাগে	</s>	(Male1/1008)
9	<s>	মানুষ মরনশীল	</s>	(Male1/1009)
10	<s>	মরিচ অনেক ঝাল	</s>	(Male1/1010)

Figure 4.2: Transcription file

### 4.1.2 Experimental Result and Analysis

The pocketsphinx framework was trained and tested with test audio files. We trained the language model on various different environments. While testing for accuracy, we provided different sentences for previous speakers to test the word error rate and sentence error rate. However, we were careful to note that the sentences string in the test phase contained the words that already exist in the vocabulary database. The results analysis would have been flawed if the string of sentences had words that the language model had not been trained in regards with the vocabulary dictionary database.

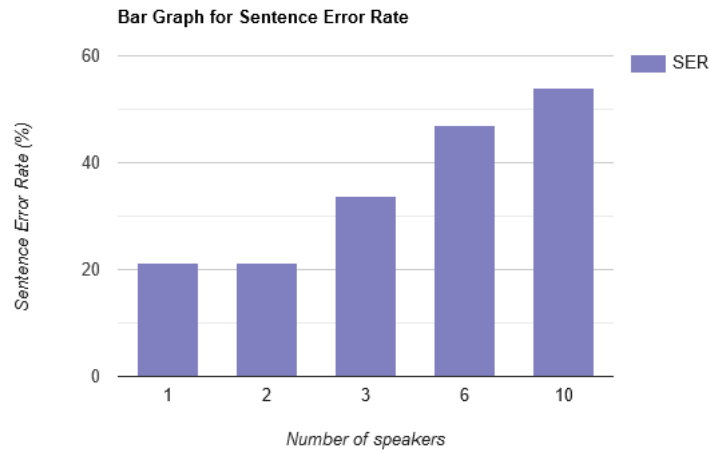
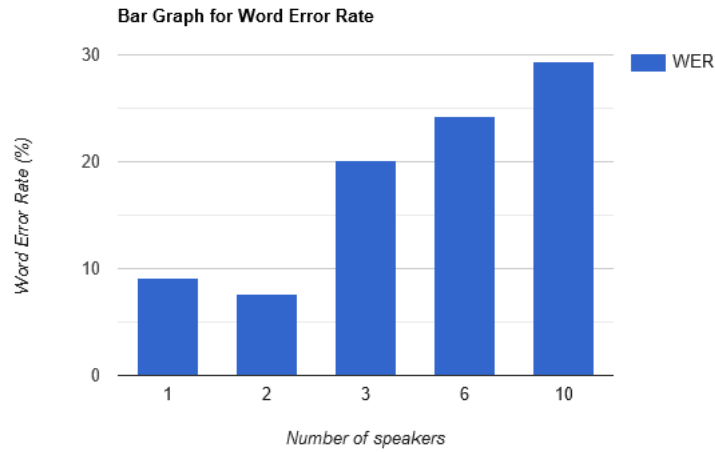
$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where

- $S$  is the number of substitutions,
- $D$  is the number of deletions,
- $I$  is the number of insertions,
- $C$  is the number of correct words,
- $N$  is the number of words in the reference ( $N=S+D+C$ )

Number Of User	Sentence Error Rate	Word Error Rate
1	21.3%	9.1%
2	21.3%	7.6%
3	33.8%	20.1%
6	46.9%	24.3%
10	54.1%	24.4%

Figure 4.3: Table with number of users to SER and WER



## 4.2 Grammatical analyzer

### 4.2.1 Data Preparation and Implementation

For training the grammatical and contextual analyzer we have collected various online resources that are available nowadays, from where Bangla Unicode sentences can be acquired. These resources include:

- Bangla online newspapers like “Prothom Alo” [11]
- Bangla blogs etc.

Special care needs to be taken when choosing well-formed sentences. In Bangla Language, various variations are found in writing linguistics in the form of “Shadhu” and “Chalito”. We are primarily focused on “Chalito”. However, sentences may contain a composite of both forms. These types of sentences may be excluded as an initial training and test phase. Additionally, spoken language may also vary greatly depending on whether the person speaks a native language or the standard Bangla

language. Next, for the test phase, we are collecting data manually from person to person to check how efficient our model is in real life.

We have collected around 1000 sentences. These sentences are manually tagged with parts of speech for training and testing. We used a data preprocessor python execution file to generate the tag sequence not to mention the word-tag vocabulary transcription file from the dataset. We have generalized the tags to noun(N), pronoun(P), adjective(J), adverb(D), verb (V)and conjunction(C).

```
J N J C J N N N C N C N
N N N N V N N C N N
J J N N C N P V C C N N N C
N C J J N N N N V P V C N D
J C J N V
N PP J J J C J V
```

Figure 4.4: Tag sequence

```
রপ্তানি J
দ্রব্য N
তাজা J
ও C
শুকনা J
ফল N
আফিম N
পশুচর্ম N
ও C
পশম N
এবং C
কাপেটি N
রাজা N
মহানন্দ N
রাজধানীতে N
তৈরি N
করেছিল V
```

Figure 4.5: Word tag vocabulary

There tag sequences is used to generate count matrix, sum matrix and the probability matrix.

The rows in the matrix represent the current states, and the columns represent the next states. The values represents the change probabilities of going from the present status to the following state.

$$P(w_1^n) \approx \prod_{i=1}^n P(w_i | w_{i-1})$$

Figure 4.6: Probability Equation

Corpus: <s>মাঠে মাঠে অনেক ফসল</s>

	<s>	</s>	মাঠে	অনেক	ফসল
<s>	0.001	0.001	1.001	0.001	0.001
</s>	0.001	0.001	0.001	0.001	0.001
মাঠে	0.001	0.001	1.001	1.001	0.001
অনেক	0.001	0.001	0.001	0.001	1.001
ফসল	0.001	1.001	0.001	0.001	0.001

Figure 4.7: Example of count matrix

Corpus: <s>মাঠে মাঠে অনেক ফসল</s>

	<s>	</s>	মাঠে	অনেক	ফসল	Sum
<s>	0.001	0.001	1.001	0.001	0.001	1.005
</s>	0.001	0.001	0.001	0.001	0.001	0.005
মাঠে	0.001	0.001	1.001	1.001	0.001	2.005
অনেক	0.001	0.001	0.001	0.001	1.001	1.005
ফসল	0.001	1.001	0.001	0.001	0.001	1.005

Figure 4.8: Example of sum matrix

$$sum(row) = \sum_{w \in V} C(w_{n-N+1}^{n-1}, w) = C(w_{n-N+1}^{n-1})$$

Corpus: <s>মাঠে মাঠে অনেক ফসল</s>

	<s>	</s>	মাঠে	অনেক	ফসল
<s>	0.0009	0.0009	1.001	0.0009	0.009
</s>	0.2	0.2	0.2	0.2	0.2
মাঠে	0.0004	0.0004	0.499	0.499	0.0004
অনেক	0.0009	0.0009	0.0009	0.996	0.0009
ফসল	0.0009	0.006	0.0009	0.0009	0.0009

Figure 4.9: Example of probability matrix

### N-gram probability

Probability of N-gram: 
$$P(w_N | w_1^{N-1}) = \frac{C(w_1^{N-1} w_N)}{C(w_1^{N-1})}$$

At the point when we execute the Viterbi calculation in the programming task, be cautious with the files, as arrangements of matrices in Python start with 0 rather than 1. Another execution explicit issue, is the point at which you duplicate numerous minuscule numbers like probabilities, this will prompt mathematical issues, so you should utilize log probabilities all things being equal, where numbers are added rather than increased.



# Chapter 5

## Challenges

Though Bangla is the 7th most spoken language in the world, there isn't that much significant work that has been done with this in terms of grammar and semantic analysis. With its complex grammatical structure and extensive vocabularies, it is very much challenging to produce a model for a complicated language and includes a complex structural grammar which we've got faced during testing our methodology.

As we've got to develop a strategy to detect the run-time error of straight forward Bangla text, there are not any such contents or corpora which are only in the simple format of Bengali grammar. Also, it's difficult to search out any standard corpus that has some run time error on the text. The use of user spoken data in research and storing user conversations necessarily raises ethical concerns. Data privacy thus takes on a central role, and the potential of abuse cannot be overestimated.

While monitoring conversations in any scenario fulfills the highest ethical standards, it could equally well be misused by anyone to monitor other people secretly. This research aims to make people's lives easier, so we have to handle sensitive data. There thus exists a proud tradition of confidentiality, whose methods can serve as blueprints for the deployment of Big Data technologies. Scaling and extending these concepts to an entirely new dimension is no mean feat, and will generate a significant amount of work for researchers, practitioners, and occupational bodies.

On a practical level, we are currently following two different privacy models. We collect user spoken data, but we refrain from leaking this data to other applications. This data is solely used for conversion to text, analysis, and further training of our model. On a different note, this paper does not automatically change user-spoken data instead suggests alternatives of words or phrases that can be used. We strive to keep the authenticity of the converted text with the highest accuracy for the best user experience.

# Chapter 6

## Conclusion

### 6.1 Conclusion

This paper introduces automated speech-to-text conversion of Bangla language with further application of grammatical and contextual analysis of the converted text. The suggested methodology stores and analyzes massive amounts of data at a low cost and without the risk of losing valuable user spoken information. The paper outlines the speech-to-text generation of Bangla language with a highly sophisticated speech recognizer, and the text is analyzed, and suggestions are displayed based on our grammatical and contextual analyzer. The proposed system outperforms traditional methods, where the users had to write down everything by hand within a limited period, both in quality and quantity. While there may be ethical concerns, these matters need to be addressed case by case with the rich data privacy policy.

### 6.2 Future Work

The algorithm is still at its primary stage. We are hoping to further develop the model so that it can perform contextual analysis and suggest alternatives for the errors it validates.

Contextual analysis: an algorithm that involves verifying whether the sentence clarifies meaning. The main objective here is to seam through the entire text and analyze whether the sentence as the whole is viable or not. There are times when a sentence may be grammatically correct but does not have any contextual meaning. We are determined to remove this uncertainty as much as possible, using the help of NLP and machine neural networks.

Suggest alternatives: for the detected grammatical and contextual errors our system will be able to suggest the corrected alternative sentences with a high accuracy rate.

# Bibliography

- [1] Lee, Akinobu Kawahara, Tatsuya Shikano, Kiyohiro. (2001). "Julius—An open-source real-time large vocabulary recognition engine. Proceedings of European Conference on Speech Communication and Technology", 1691-1694.
- [2] Lamere, Paul Kwok, Philip Walker, William Gouvêa, Evandro Singh, Rita Raj, Bhiksha Wolf, Peter, "Design of the CMU sphinx-4 decoder".
- [3] Chowdhury S. A., Khan P. M., "Implementation of Speech Recognition System for Bangla", School of Engineering and Computer Science BRAC University, Bangladesh, August 2010
- [4] Nahid M. M. H., Islam M. A. and Islam M. S., "A Noble Approach for Recognizing Bangla Real Number Automatically Using CMU Sphinx"
- [5] Arif H., "Bangla Speech to Text Conversion Using CMU Sphinx", Department of Computer Science and Engineering, BRAC University, Bangladesh.
- [6] Sultana S., Das P. K., Akhand M. A. H. , Rahman M. M. H. "Bangla Speech-to-Text conversion using SAPI", NotreDame University Bangladesh, Khulna University of Engineering and Technology Bangladesh, Khulna University of Engineering and Technology Bangladesh, King Faisal University Bangladesh, July 2012.
- [7] Hasan K. M. A., Al-Mahmud, Mondal A, Saha A., "RECOGNIZING BANGLA GRAMMAR USING PREDICTIVE PARSER", Department of Computer Science and Engineering, Khulna University of Engineering and Technology, Bangladesh, (IJCSIT) Vol 3, No 6, Dec 2011
- [8] Hasan K. M. A., Hozaiifa M. and Dutta S., "Detection of Semantic Errors from Simple Bangla Sentences", Computer Science and Engineering Department Khulna University of Engineering Technology Khulna 9203, Bangladesh, December 2014.
- [9] Haque M. M., Habib M. T., Rahman M. M., "AUTOMATED WORD PREDICTION IN BANGLA LANGUAGE USING STOCHASTIC LANGUAGE MODELS", Dept. of Electrical and Computer Engineering, North South University, Bangladesh, Dept. of Computer Science and Engineering, Daffodil International University, Bangladesh, Dept. of Computer Science and Engineering, Prime University, Bangladesh, International Journal in Foundations of Computer Science Technology (IJFCST) Vol.5, No.6, November 2015

- [10] Kundu B, Choudhury S. K., Chakraborti S., “NLG Approach for Bangla Grammatical Error Correction”, Language Technology, ICT Services Centre for Development of Advanced Computing Kolkata-700091, Department of Computer Science Engineering Indian Institute of Technology, Chennai-600036, India.
- [11] Prothom Alo, Tuesday, 6 October 2020, <https://www.prothomalo.com/>
- [12] Poka, Ahmed H., 1993, Dipu Number Two, Muhammed Zafar Iqbal, 2013
- [13] Arora Y., Kaur J., Gupta A., “Grammatical Error Correction using Neural Networks”, ICME and Department of Electrical Engineering Stanford University, United States of America.
- [14] Kabir H., Ahmed R., Nasib A. U., “Real Time Bengali Speech to Text Conversion using CMU Sphinx”, School of Engineering and Computer Science,Brac University, Dhaka,Bangladesh.
- [15] Ghadage, Y. and Shelke, S. (2016). ”Speech to text conversion for multilingual languages” - IEEE Conference Publication.
- [16] Shmyrev, N. (2019). Building a phonetic dictionary. [online] CMUSphinx Open Source Speech Recognition. Available at: <https://cmusphinx.github.io/wiki/tutorialdict/> [Accessed 30 Jul. 2019].
- [17] Sultana,Palit R., Rajesh. (2014). A survey on Bengali speech-to-text recognition techniques. 2014 9th International Forum on Strategic Technology, IFOST 2014. 26-29. 10.1109/IFOST.2014.6991064.
- [18] A. Bapat and L. Nagalkar, ”Phonetic Speech Analysis for Speech to Text Conversion - IEEE Conference Publication”, Ieeexplore.ieee.org, 2008 [Accessed: 06- Mar- 2009].
- [19] Shmyrev, N. (2019). Basic concepts of speech recognition. [online] CMUSphinx Open Source Speech Recognition. Available at: <https://cmusphinx.github.io/wiki/tutorialconcepts/> [Accessed 1 Aug. 2019].
- [20] J. L. Nusrat, N. E. Qamrun, M. Ghulam, Dr. N. H. Mohammad, Prof. Dr. M. R. Rahman, “Performance Evaluation of Bangla Word Recognition Using Different Acoustic Features,” IJCSNS International Journal of Computer Science and Network Security, vol. 10, No. 9, September 2010.
- [21] B. Zhao and X.-H. Tan, “Oral English training system based on speech recognition technology,” Journal of Computer Applications, vol. 29, no. 3, pp. 761–763, 2009.
- [22] Maegawa, S. (2012). SPEECH RECOGNITION METHOD, SPEECH RECOGNITION SYSTEM AND SERVER THEREOF. The Journal of the Acoustical Society of America, 131(6), p.4868
- [23] Sean R. Eddy, “Multiple alignment using hidden Markov models,” Dept. of Genetics, Washington University School of Medicine, 1995.
- [24] Sorensen, J., and Allauzen, C., “Unary data structures for Language Models”, INTERSPEECH 2011.