Bangla Speech to Text Conversion Using CMU Sphinx

by

Israt Jerin Bristy
15301006
Nadim Imtiaz Shakil
15301037
Tesnim Musavee
15101110
Akibur Rahman Choton
15301102

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
August 2019

Declaration

It is hereby declared that

1. The thesis submitted is 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:

Israt Jerin Bristy
15301006

Nadim Imtiaz Shakil
15301037

Tesnim Musavee
15101110

Akibur Rahman Choton
15301102

Approval

The thesis titled Bangla Speech to Text Conversion Using CMU Sphinx by

1. Israt Jerin Bristy (15301006)
2. Nadim Imtiaz Shakil (15301037)
3. Tesnim Musavee (15101110)
4. Akibur Rahman Choton(15301102)

Of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 7, 2019.

Examining Committee:

Supervisor:
(Member)

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Jia Uddin
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbub Alam Majumdar
Professor
Department of Computer Science and Engineering
Brac University

Abstract

Speech is the most normal type of communication and association between people while content (text) and images are the most basic types of exchange in the computer system. Therefore, enthusiasm in regards to transformation between speech and text is expanding day by day for integrating the human-computer relation. Understanding speech for a human is not a challenge but for a machine it is a big deal because a machine does not catch expression or human nature. For the conversion of speech into text, this proposed model requires the usage of the open sourced framework Sphinx 4 which is written in Java. For the proposed system, it requires certain steps which are training an acoustic model, creating a language model and building a dictionary with CMUSphinx. For training, the audio files were recorded by 8 speakers both male and female for more accuracy. Among them, 6 speakers recorded each word 3 times. To test the accuracy, we took audio recordings from 2 speakers among them one speaker is unknown to the system. After testing, we got the accuracy around 59.01%. For known speakers we got 78.57% accuracy. We gave audio files as input only to check accuracy as our main purpose was to make a system which works in real time. In our system, user can speak in real time and the system converts it into text.

Keywords: Bangla; Voice Recognition; CMUSphinx; Acoustic model and Language model

Acknowledgement

This is the work of Israt Jerin Bristy, Nadim Imtiaz Shakil, Tesnim Musavee and Akibur Rahman Choton- students of the CSE department of BRAC University. The document has been prepared as an effort to organize the knowledge acquired by us about the thesis. All thanks to Almighty who provided us guidance and capabilities to complete this research. We would like to express our sincere gratitude to our supervisor for his invaluable guidance, comments and suggestions during the course of our thesis. The completion of this study could not have been possible without his continuous support regarding the study. We are also grateful to the faculty members of the department of Computer Science and Engineering of BRAC University, who have always been helpful and kind towards us in this whole study period. We take this opportunity to express gratitude to all of the Department faculty members for their help and support. We also thank our parents for their encouragement and support.

Table of Contents

List of Figures

Chapter 1

Introduction

In the era of information and communication technology, the applications of speech recognition system and speech to text conversion are increasing day by day. However, there has not been enough research or applications regarding Bangla speech to text conversion. Bengali or Bangla, is the official and most widely spoken in Bangladesh and also spoken in different parts of the Brahmaputra, Andaman and Nicobar Islands in the Bay of Bengal including Jharkhand, Bihar, Orissa, Meghalaya and Mizoram with approximately 250-300 million total speakers worldwide. There is no other language in the world for which people had gone through any kind of language movement. The movement of 1952 marked Bangla as the International Mother-Language day. Hence, 21st February is celebrated as the International Mother Language day worldwide. That is why we tried to do our thesis on Bangla speech to text conversion.

1.1 Motivation

Working on the topic based on machine learning, artificial intelligence, big data, etc is challenging and also need the patience to work. But, here our motivation to work on a bit different topic than usual is to take more challenges and obviously overcoming those.

At first, our team decided to work on 'Predicting Accurate Cricket Match Score' using machine learning. But one of us came up with this interesting topic of Bangla Speech to text conversion. Why did we choose this topic whereas machine learning works are famous and available in today's market? We wanted to take challenges and come up with solutions. Yes, we almost could do this. There were not that much of resources we could find over the internet. Besides, this project pushed us 3 or 4 times to do the same things since our computer was not helping us properly.

Moreover, we found speech recognition very interesting to work with since we found a very new framework called CMU Sphinx4 which offers to work with Java, C, C++, etc. So we had choices for programming language to work with. Also, we found there were thesis works on speech recognition based on TensorFlow framework but we choose CMUSphinx over that. So that we can get a better result than previous works on CMUSphinx, so that can improve

the system, we can develop a new algorithm which was tough for us but tried our best.

This research helped us to know the variety of new kinds of stuff, technologies, way of solving problems which were our main goal. Our motivation was to develop a better system which we tried our best.

1.2 The Goal of the Thesis

The aim of the project is not only solving the difficulty of typing in Bangla but also giving options for practical use. First of all, the proposed system would be very helpful for people who are technically challenged or cannot type. This can be a game-changing tool for the illiterate people as their medium of communication will improve. Furthermore, the system also aims to help deaf people by giving assistance to the academic sector. Moreover, our proposed model can also create a great role in the journalism sector. From this application, a journalist does not need to listen over and over again to take notes. So, the main goal of our project is to support the Bangla language by adding new applications for this language.

1.3 Contribution

There had not been many works on Bengali speech to text conversion and also in real-time recognition. That was the reason we attempted to build a real-time Bangla Speech to Text Conversion system. For the conversion of speech into text, this proposed model was made by Sphinx 4 .The Sphinx4 was an open source framework. Moreover, Sphinx4 was written in Java. With the help of CMU Sphinx, we trained an acoustic model. We built a dictionary with CMU Sphinx. We also made a language model which was necessary for the proposed system. Furthermore, we also created phone-set, File id and Transcription file to train and test our system. In our dataset it contains 101 words and recorded by 8 speakers including both male and female.

Our proposed system works in every platform. Moreover, this system runs in real time which means it takes input and gives output at the same time. The system also works for continuous speech and we added 101 Bengali words for conversion.

Chapter 2

Background Study and Literature review

2.1 Theoretical Discussion

For the required project, at first, we want to make a speech recognition system. Speech recognition (SR) is that the method of changing associate degree acoustic signal, captured by a electro-acoustic transducer or a microphone, to a collection of words. to grasp the speech recognition system properly, First we want to understand the structure of speech. Speech may be a advanced development. individuals seldom perceive however is it created and perceived. The naive perception is commonly that speech is made with words and every word consists of phones. the fact is sadly terribly totally different. Speech may be a dynamic method while not clearly distinguished elements. It's perpetually helpful to induce a sound editor and appearance into the recording of the speech and hear it. Here is for example the speech recording in an audio editor.

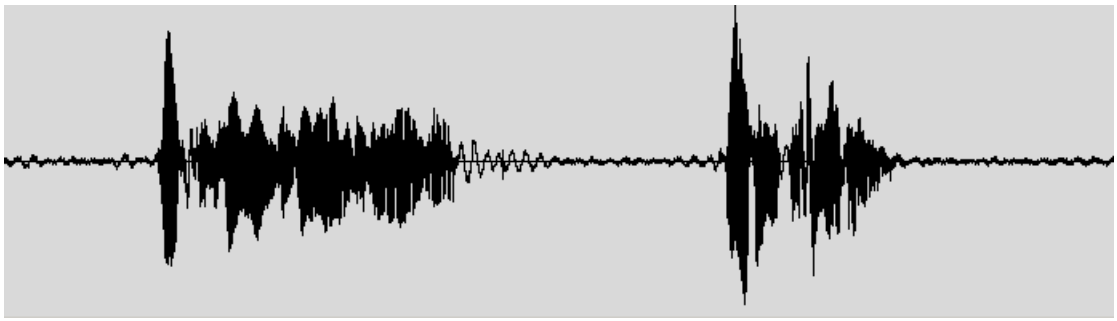

Figure 2.1: Speech Recording in an Audio Editor

Speech is a natural action that is the expression the flexibility to specific thoughts, feelings and emotions by articulate sounds or verbal words. therefore speech is that the communication or expression of thoughts in spoken words in numerous languages that is completely controlled by human vocal. Every language has its own constants, vowel and acts. In current follow Speech may be a continuous audio stream wherever one will outline a lot of or less categories of sounds or phones. The acoustic properties of a wave shape of phone will vary on several factors. Parts of phones between 2 consecutive phones is thought as diphone. Besides, totally different sub-states of a phone is thought as subphonetic units. There area

unit 3 states of phones that area unit relying on preceding phone, the center half is stable and also the next half depends on the next phone. Triphones or quinphones area unit thought-about in context. CMUSphinx used 4000 distinct short sound detectors to compose triphones. These area unit referred to as senones. What is more, phones build sub-word units as example syllables that area unit associated with intonational contour. It is necessary to say concerning words in speech recognition as a result of they limit combinations of phones considerably. Words and different non-linguistic sounds area unit referred to as fillers that are kind of utterances and conjointly separate chunks of audio between pauses.

All recent descriptions of speech area unit to some extent probabilistic. which means that there are no sure boundaries between units, or between words. Speech recognition systems can be classified into the subsequent teams counting on input Patterns.

1. Speaker Dependent vs. Speaker Independent: within the speaker dependent recognition system, speech of a selected speaker is employed to coach the popularity system to recognize his/her speech. On the opposite hand, speaker independent recognition system will recognize anyone's voice because it is not speaker sensitive

2. Isolated vs. Continuous Speech Recognition: In Associate in Nursing isolated speech recognition system, the speaker delivers precise pause between 2 words however there's no clear pause between the spoken words for continuous speech. Continuous speech is the real life conversations, and never-ending speech recognition system is appropriate for sensible or real-world usage.

3. Keyword based vs. Sub-word Based: The speech recognizer will be trained to acknowledge complete words that area unit referred to as keyword dependent recognizer. Syllables or phonemes may be recognized by coaching the recognizer that area unit categorized as sub-word dependent recognizer.

Figure 2.2: Overview of Speech Recognition System [1]

In our thesis, we build a speaker independent recognition system for Continuous Speech. To develop speech recognition system we used CMUSphinx which is a very popular speaker independent system amongst researchers working on Automatic Speech Recognition (ASR). With the help of CMUSphinx, we trained an acoustic model, build a language model and a dictionary which are necessary for the conversion of speech to text.

2.2 Literature Review

We have gone through some research, journals, conference paper and project papers to understand the concept and process for Bangla speech to text conversion. Among the few types

of research, research on Speech-to-Text is rare compared with Text-to-Speech. Some of the existing papers and related works regarding speech to text conversion are as follows.

We have gone through a paper that is known as Implementation of speech recognition system for Bangla [1]. For the implementation to setup the task they used Perl and C compiler. The project has been entirely setup in Linux platform, therefore to examine whether or not each Perl and C compiler are gift. they had gone through with varied experiments and so the take a look at lead to four experiments there are shown the small print and results clearly. This implementation of Speech Recognition System has been engineered on little information, domain based mostly and trained with solely half dozen speakers. their intention was to integrate the system with Interactive voice response.

There is a paper named Bangla Speech-to-Text Conversion using SAPI [2] relates to our thesis. In this paper, they build a speech to text conversion for Bangla language using Speech applications program Interface (SAPI) that is developed by Microsoft Corporation. In their paper, they used SAPI to match pronunciation from continuous Bangla speech in precompiled synchronic linguistics file of SAPI and SAPI came Bangla words in English character if matches occur. The words are then accustomed fetch Bangla words from the information and come back words in true Bangla characters and to complete the sentences. Moreover, they additionally with success overcome pronunciation variation yet as tone variation in language.

In the paper named South Dravidian Speech to Text Conversion using CMUSphinx [3], the authors projected a completely unique South Dravidian automatic Speech to Text conversion System (ASTC). In their project, they train and take a look at, the Speech process System using CMUSphinx framework that is dynamic in nature with support for different languages beside English. what is more, they trained the Acoustic model for South Dravidian speech with one thousand general spoken sentences and tested one hundred fifty sentences. With the assistance of utilizing options in CMUSphinx, they additionally build their system. during this paper, South Dravidian sentences with four to ten word length are researched, and that they experimented speech to text conversion system for the South Dravidian Language with restricted speech corpus of thousand sentences.

There we got another paper which is Real Time Bengali Speech to Text Conversion using CMUSphinx [4]. In the paper they used framework is Sphinx 4. Free digital audio workstation has been used to collect recording data. 10 individual speakers consisting of both males and females are being recorded with good accuracy. The system has been divided into two different segments named 'Training' and 'Testing'. They have created Two Algorithm for this system. Audacity is used for audio correction and manipulation. The ideal idea was to about 200 different speakers with around 10 minutes of data per speaker to get the optimum accuracy which the sphinx 4 documentation says would be the optimum amount for a well-trained model. Their aim is to potentially be applicable to iOS, Android and Windows phone and also UI for the interface of usage for the system which device they can have in their pocket.

Dana Dann'ells [5] presented a work supported combinations of lexical, semantic, syntac-

tical and acoustic data for measure disfluency rates, and therefore the knowledge sample. supported empirical proof and theoretical findings, computing linguistic associate degreed non-linguistic options are a robust approach in association with alternative attributes to see whether or not an vocalization is disfluent and its signals are a psychological feature load. Neema Mishra et.al (2013) [6] combined strategies like feature extraction, acoustic modelling and language modelling in Sphinx methodology for distinctive disfluencies within the Hindi language. The system was ready to acknowledge Hindi audio with success. Satori et.al (2007) [7] projected a way to spot disfluency in Arabic speech. They used Sphinx methodology having associate degree acoustic model, Language model and phonetic dictionary.

Yang Liu et.al (2003) [8] projected a way to seek out disfluencies in an exceedingly speech, they used acoustic-prosodic options, language model (LM), part-of-speech (POS) based mostly lumen, and rule-based data. They trained the system with switchboard-I corpus with speech conversations and tagged the disfluencies. They obtained the accuracy in distinctive disfluencies with a exactitude rate of 68% and a recall rate of 46%. Matthias Honal et.al (2003) [9] projected a way for locating disfluencies in an exceedingly speech. applied math artificial intelligence approach is being employed during this method that the blatant channel adds noise to its input sentence for disfluencies and creates language as output. They experimented with English and Chinese language. For English, finding a disfluency rate was 72% recall and 92% exactitude. They used Mandarin dialect knowledge.

Yulia Tsvetkov et.al (2013) [10] proposed a way to handle disfluencies, word phrases and self-interruption points in Yue dialect colloquial speech knowledge. The strategies used are Word Fragments Identification and Word Fragments Modelling in conjunction with Annotation of word fragments. The accomplishment of accuracy is 88% for automatic detection of disfluencies.

In this paper [11], the authors described their research on challenges and techniques of different speech to text algorithm. They discussed about some challenges such as Utterance approach, Utterance Style, Types of speaker model, Vocabulary, Channel Variability. In these challenges the author explained about whether the words are isolated or connected words. Moreover, the paper says about utterance style like continuous speech and spontaneous speech. Different types of vocabulary were described, for example small, medium and large vocabulary. In techniques section they preferred to discuss about different types of feature extraction techniques such as MFCC, LPC, LPCC, PLP, LDA, DWT, RASTA-PLP, PCA. However, the author mentioned three types of speech recognition which are Acoustic Phonetic Approach, Pattern Recognition Approach and Artificial Intelligence Approach. To conclude with, the paper offers the fundamentals of speech recognition system along with different approaches which are available for the system.

This paper includes speech recognition system for Bangla real number where the authors used CMUSphinx4 [12]. The authors from SUST here emphasized on real numbers and they tried to implement the system so that the system could provide accurate numbers as output after speech is delivered. They used almost 3207 sentences and 115 words for the system which took 3.79 hours of recordings. Also in training phase they used sampling rate of 16000

Hz, channel was mono. In case of the accuracy they did it kind of manually, such as they experimented on 100 words and they found 85 words accurately and the rest 15 was wrong. Overall, after their training and testing they could secure 75% accuracy according to their paper.

In this paper, the authors from KSA and Malaysia took step to use CMUSphinx4 to recite the holy Quran and fault tolerance [13]. They tried to train the system with data which were very challenging. In order to do that, they developed a tool to generate all the required files. Since the holy Quran is huge in case of number of sentences or words, it was really time consuming task and hard as well. For the system, 80% of audio files were selected for training and 20% for the testing purpose. The system took about 50 minutes of training and 3030 seconds for different training configurations. They tried to add sufficient pauses in between Ayah and the words to make the system and Ayah accurate enough. To conclude with, though training the system was tedious enough but they tried to make it more accurate and worked on fault tolerance in different cases which helped them to make it accurate enough.

There is a paper that was published in IEEE named Speech to text conversion for multilingual languages that is predicated on information in speech language [14]. The most objective of this method is all regarding extract characterize and acknowledge the data regarding speech. The system is enforced using Mel-Frequency Cepstral Co-efficient, feature extraction technique and Minimum Distance Classifier, Support Vector Machine strategies for speech classification. during this system speech utterances square measure pre-recorded and data-base is especially divided into 2 elements that square measure testing and coaching. The system was developed in MATLAB setting that is being mentioned. During this system they principally used 2 languages and conjointly the mixture of these 2 languages. These two languages square measure English and Marathi and conjointly the combination of English and Marathi. They had a thought to increase the system for different regional languages and extend towards the 64000 time connected word speech recognition for various languages.

We found a paper named hierarchical bayesian Language Models for colloquial Speech Recognition supported a statistic previous known as Pitman-Yor process[15]. The main goal of the language model is smoothing, embedding the power-law distribution for language. within the system they planned the applying of hierarchical Pitman–Yor method language models on a large-vocabulary ASR system for colloquial speech, victimisation moderately massive corpora and conjointly regarding comprehensive experimental results on multiparty colloquial meeting corpora, with the observation that the HPYLM outperforms each the IKNLM and also the MKNLM.

The paper A SVM primarily based speech to text convertor for Turkish language conjointly a couple of system that is on speech to text converter[16]. Here Mel Frequency Cepstral Coefficients (MFCC) has been applied to extract options of Turkish speech and SVM primarily based classifier has been wont to classify the phonemes. Within the system, a replacement Text Comparison algorithmic program is planned that conjointly used speech sound sequence to live similarity in word similarity measuring.

Phonetic Speech Analysis for Speech to Text Conversion is a paper where we tend to found the system is regarding generating phonetic codes of the verbalised speech in training-less, human freelance manner[17]. The system work is target-hunting by the operating of ear in response to audio signals. The system explains and proves the scientific arrangement of the Devnagari script and tries to phase speech into speech sounds and establish the phoneme victimisation easy operations like differentiation. however the information employed in coming up with the system isn't utterly noiseless; the system do not deals in removing noise however operates input file directly assumptive its noise. It is expected the length of silence before a plosive vocalization are a decent parameter.

We go through a paper named Applying Speech-to-Text Recognition and Computer-Aided Translation for Supporting Multi-lingual Communications in society Learning Project that may be a system to support multi-lingual communications students collaborating in society learning project[18]. The participants were engaged in interactions and knowledge exchanges and communications were allotted in their native languages on social communication platforms. Moreover, they aimed to look at accuracy rates of processes related to STR and CAT for various languages throughout multi-lingual communications in their society learning project. Their results showed that very cheap accuracy rate was for Mongolian and Filipino and also the highest was for Spanish, Russian, and French.

Chapter 3

Proposed Model

3.1 Algorithm

3.1.1 Hidden Markov Model

Hidden Markov Model (HMM) have demonstrated to be one of the most generally utilized devices for learning probabilistic models of time arrangement information. In a HMM, data about the past is passed on through a solitary discrete variable—the hidden state. We talk about a speculation of HMMs in which this state is considered into various state factors and is consequently spoken to in a dispersed way. We depict a careful calculation for inducing the back probabilities of the shrouded state factors given the perceptions, and relate it to the forward–in reverse calculation for HMMs and to calculations for progressively broad graphical models. Because of the combinatorial idea of the shrouded state portrayal, this careful calculation is unmanageable. As in other immovable frameworks, estimated induction can be completed utilizing Gibbs testing or variational methods. Inside the variational system, we present an organized guess where the state factors are decoupled, yielding a tractable calculation for learning the parameters of the model. Observational correlations propose that these approximations are effective and give precise options in contrast to the accurate strategies. Finally, we utilize the organized guess to demonstrate Bach's chorales and demonstrate that factorial HMMs can catch measurable structure in this data set which an unconstrained HMM cannot.

However, CMUSphinx4 framework, which we used for our system was being developed based on this Hidden Markov Model (HMM). It is viewed as the 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. The Viterbi calculation works related to the Hidden Markov Model to locate the doubtless succession of concealed states.

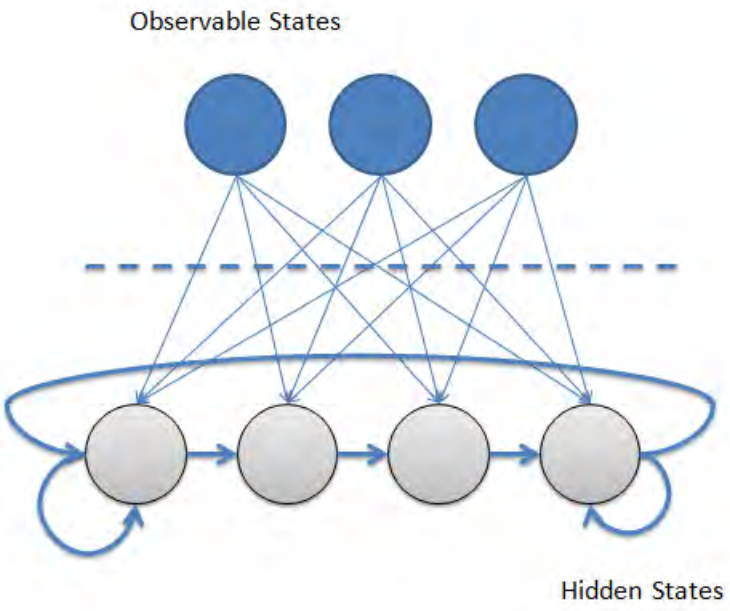

Figure 3.1: Hidden Markov Model

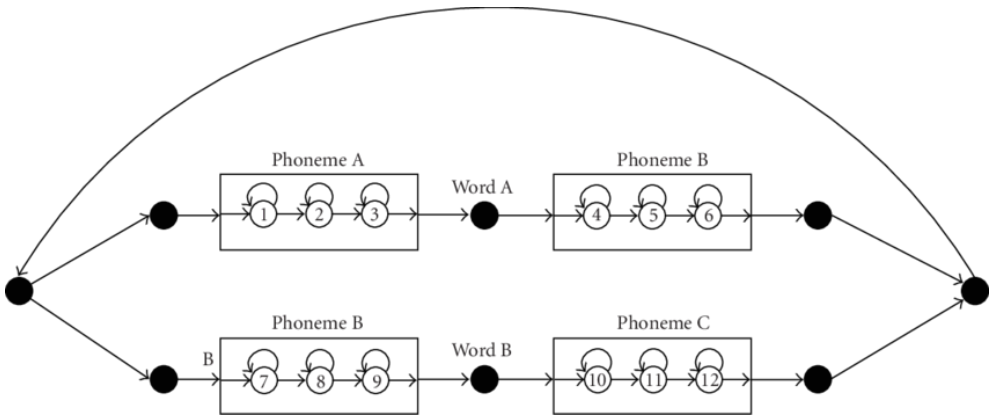

Figure 3.2: Hidden Markov Model in CMUSphinx

3.1.2 Word Matching Algorithm

Our team developed this algorithm based on how we developed the system. We explained how the words were being converted from speech to text and how the voice recordings got converted.

At first we needed to collect recordings of different speakers for unique words. Then according to those unique words we added phone against those specific Bengali words in dictionary

file. Besides, we also added unique phones in phone file where all the phones were used in dictionary file. To make understanding for the system, we needed to create a bridge, here other_train.fileids file worked as like as the bridge, so that the system could recognize the words for individual speaker and can match it in the testing phase.
Moreover, after completing all of these steps, the most important step, this was to add Bengali words against each filename in <s> </s>.

Algorithm 1 Word Matching algorithm

```
for <each Benagali Word> do
    <Creating Phones for each Bangla Words>
    <Adding the Phones beside the Bengali words in other.dic file/list>
end for
for <each Bengali Word> do
    for <each Phone> do
        if <Phone don't exist > then <Adding the Phones into the other.phone file/list>
        end if
    end for
end for
for <each File> do
    <Adding file as folderName/fileName (Such as speaker1/ami)>
    <Creating train.fileids files>
end for
for <each Recording> do
    <Adding expected Bangla word against recorded filename>
    <Adding the Bangla words into <s> </s> accordingly>
end for
```

3.2 System Design

There are many language models database that support English which enhances the accuracy of English speech recognition. But a few work has been done on Bangla which makes it hard to build an acoustic model for Bangla. For the proposed model, we created a speech recognition system and train it with our recorded data. Configuring the system to work for Bangla language needs a series of steps with CMUSphinx.

3.2.1 Tools We Used

As we already mentioned that we used CMUSphinx framework for the speech recognition. This toolkit is a leading speech recognition toolkit with various tools used to build speech applications. To set up the environment for CMUSphinx in our system, first we installed python and perl. CMUSphinx contains a number of packages for different tasks and applications. For the proposed system, we used latest version of the followings.

- Pocketsphinx: It's a lightweight recognizer library written in C.
- Sphinxbase: It supports library for Pocketsphinx
- Sphinxtrain: It's an acoustic model training tools[19]

For this tools, we needed to install visual studio and download sphinx from their site. Then we extracted them and built the batch file. The CMUSphinx framework relies on Hidden Markov Model or HMM that has three key models for detecting any new language which are the acoustic model (AM), language model (LM), and phonetic dictionary file [20]. For the system, we created an automatic speech recognition (ASR) system which have this models. The decoder in this system is the software that actually performs the conversion from speech to text. On our thesis, we are using PocketSphinx for that. Figure 3.3 shows the automatic speech recognition system.

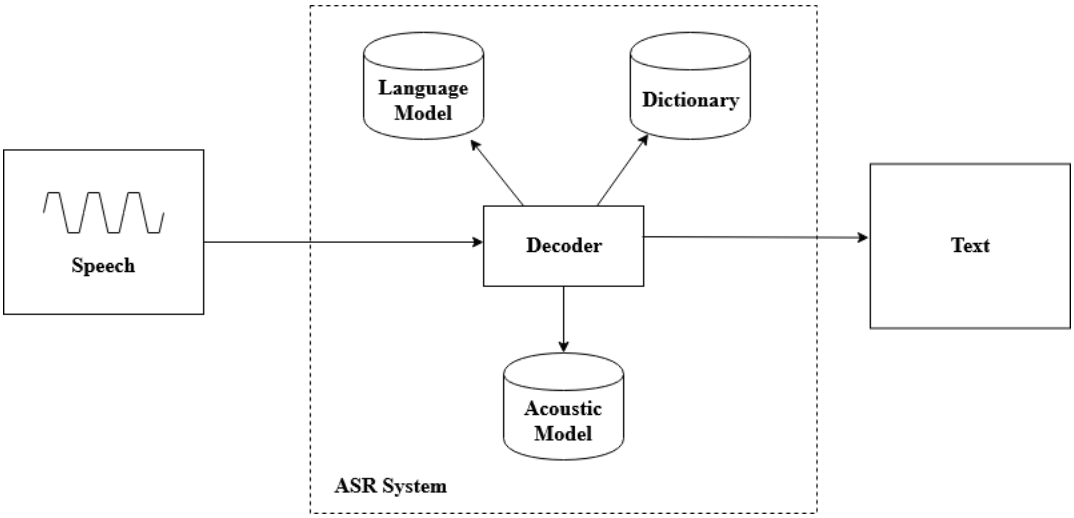

Figure 3.3: Automatic Speech Recognition System

3.2.2 Language Model

The first component of speech recognition system is the Language Model. The language model is an important component because it tells the decoder which sequences of words are possible to recognize. Language Model describes the probability of occurrence of a given word based on n-1 previous words (n-gram model, with typically n=3). For the proposed model, we build a statistical language model which represents the structural constraints available in the language to generate the probabilities. It also specifies what are the valid words in the language and their arrival sequence in the speech data.

As our data set is limited, we used an online web service which is

http://www.speech.cs.cmu.edu/tools/lmtool-new.html.

In this website, we uploaded a corpus.txt file containing 102 Bengali words and created a language model. The following figure shows a sample of our language model with its statistical values on the third gram.
In this figure the first column shows the probability in terms of logarithm (base 10) of conditional probability 'p' thus giving a negative result. The second column shows the word we recorded.

Probability	Words
-0.3010	অন্ধকার
-0.3010	আকাশ
-0.3010	আছ
-0.3010	আট
-0.3010	আমরা
-0.3010	আমাদের
-0.3010	আমার
-0.3010	আমি
-0.3010	আলো
-0.3010	এক

Figure 3.4: Language Model with the Probability for Each Word

3.2.3 Dictionary File

The second component is the Dictionary, which contains of the words and their pronunciation phoneme. It provides the system with a mapping of vocabulary words to sequences of phonemes [20]. At first the input audio consults the Acoustic Model to determine what the individual phones would be. After that it cross references the detected phones with the dictionary file to see which sequences best match the word list. After building the language model a dictionary file created which has 102 words in Bengali language. For our thesis the file is named as [other.dic]. In that file, we manually added all the phonemes for each word.

In this dictionary file each new line contains a unique Bengali word followed by a tab in the actual text file and then its respective phone sets sequence separated by spaces. Figure 3.5 showing a dictionary file contains the words and phonemes for each word.

If the input file does not find a match in dictionary, it will search in the language model to determine the closest match. This dictionary file must only contain words that exist in the transcript and vice-versa thus it is very error prone to create manually and also very hectic to get right with each training phase. There is also a file named filler file to handle silence in recordings.

Words	Phone Set
অন্ধকার	O NDHO KA R
আছ	A CHO
আট	A T
আমরা	A M RA
আমাদের	A MA DE R
আমার	A MA R
আমি	A MI
আলো	A LO
এক	E K
করছি	KO R CHI

Figure 3.5: Dictionary File Containing Phonemes for each Word

3.2.4 Acoustic Model

The third component of the linguist is the Acoustic Model, which describes sounds of the language. It provides a mapping between a unit of speech and an HMM. Machine learning techniques are usually applied to train hidden Markov Models (HMM) and Gaussian mixture models (GMM) to be store on the AM. For Bengali language there is no existing acoustic model that is why we had to build and train an acoustic model. After creating language model, dictionary, fileids and transcription, we wrote a python command.

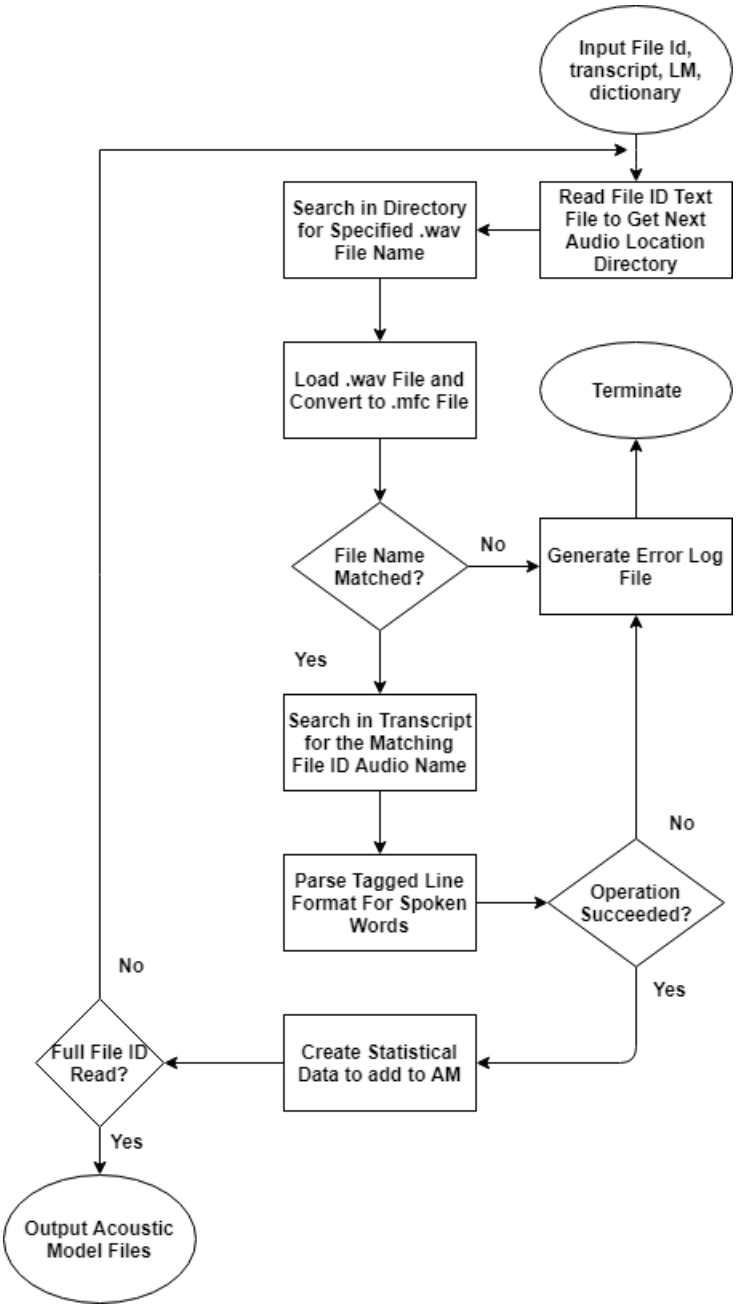

Figure 3.6: Acoustic Model Generation Flowchart [4]

3.2.5 Word Detection Model

Figure 3.7 shows the various steps of how the detection model is working starting from voice Input to text output. At first the input voice will be dissected at silence to form individual word. Then detected Phone sets for each sample will cross reference from acoustic model, the language model and dictionary and print the matched word. To do that, we took all possible combinations of words and try to match them with the audio. We chose the best matching combination. This process repeats iteratively until the full duration of the given input waveform has been converted to text.

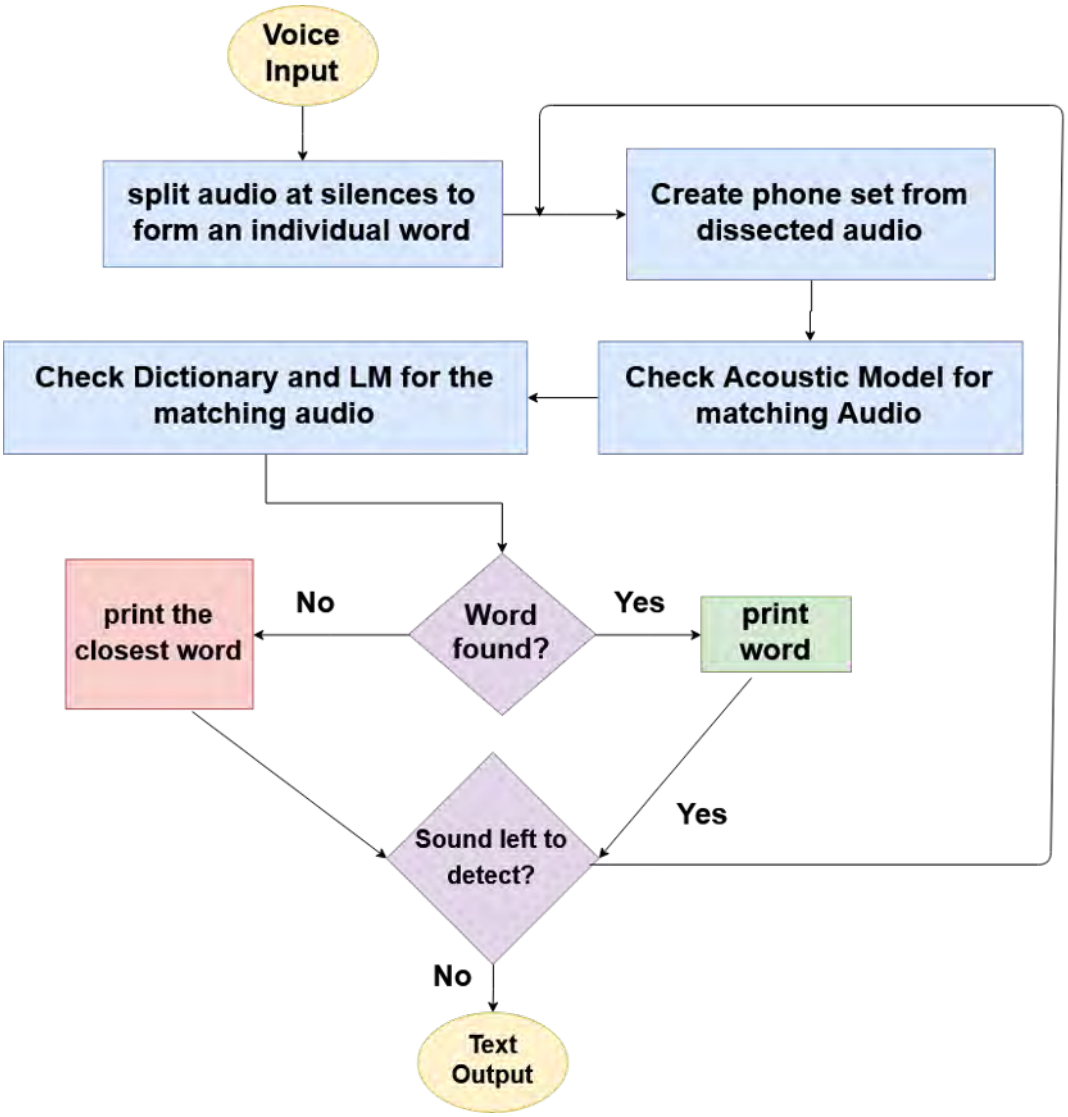

Figure 3.7: Word Detection Model

Chapter 4

Implementation and Result Analysis

4.1 Implementation

4.1.1 Data Collection

To build a general speech recognizer, lots of data would have been required which was not easy to collect in this short period of time. For our thesis, we recorded 101 words and listed these 101 words in corpus.txt file in Bengali language. The words we selected was for generating random sentences. Additionally, total 8 speakers recorded each word. For better accuracy, among these speakers 6 of them recorded each word 3 times and 2 of them recorded each word one time.

4.1.2 Phone Set

This is a file containing a list of every unique phone used for every word in the dictionary. These phones are the fundamental building blocks of syllables of words in a recording. They are written one phone per line, often as either two characters or a single character and their combination form the pronunciation of the full 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 very fundamental file to get right as it drastically affects the final accuracy of detection. It is not always clear what the individual phones of a word may sound like in broken form and sometimes may not be possible to be automated by the dictionary tool as it can lead to very bad WER (Word Error Rate). So, often the correct conversion of words to phonemes can be an iterative process with certain languages, especially ones as complex as Bangla which contains symbols used as conjunctions with base characters in order to modify the utterance. Figure 4.1 shows a simplified visual process of the various phases of conversion from speech to waveform to phone sets to "Banglish" which is the phonetic form of Bengali written in English and then finally Bengali text. For example, the Bengali word "" can be written as "Amra" in "Banglish" form.

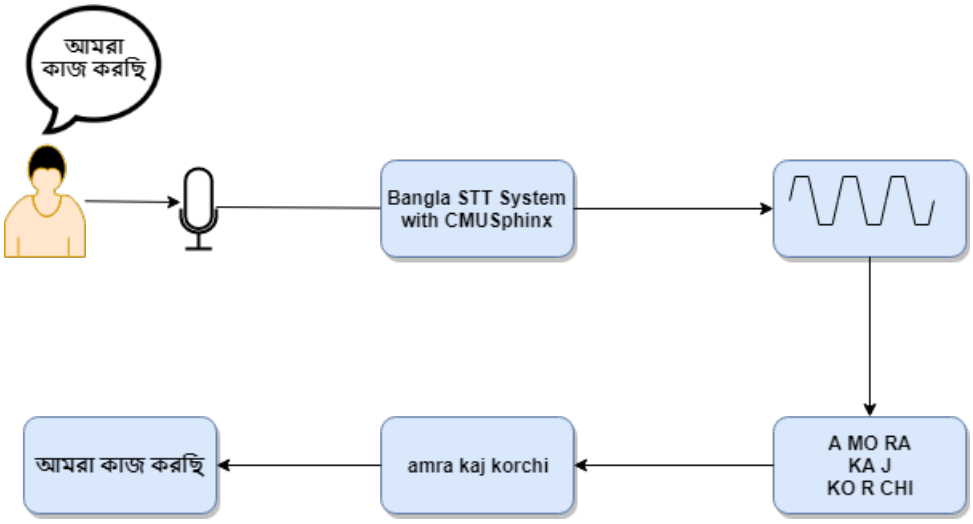

Figure 4.1: Vocal to Phone Set Conversion

4.1.3 File ID

This file handles keeping track of all the audio files being used for training and their respective directories inside the training folder. The transcript file uses this to map its sentences with the provided audio WAVE file. During the training process it is also used to keep track of all the files to convert into a special format called MFC (Mel-Frequency Cepstrum) which represents the short-term power spectrum of a sound that is required for training efficiency. To create this file dynamically we wrote a script in python. Figure 4.2 shows the structure of train.fileids file. In this file it shows the directories of every audio file. Here, Speaker1 is the folder containing all the recordings of speaker 1. For our convenience, we named each audio file by its actual word recorded in it so that we can easily understand what is recorded in that file.

```
speaker1/aat
speaker1/acho
speaker1/alo
speaker1/amader
speaker1/amar
speaker1/ami
speaker1/amra
speaker1/batash
speaker1/beton
speaker1/bhai
speaker1/bhat
```

Figure 4.2: Audio File Directories From train.fileids

4.1.4 Transcription File

A transcript is required to represent what the speakers area unit expression within the audio file. thus in a very file the dialogue of the speaker noted precisely the same precise method it's been recorded, with starting tag , ending tag followed by the file name that represent the auditory communication.This file is understood as transcription file and primarily there are 2 styles of transcription file.One of them is employed to train the system and another is to test. The transcript file is required during the training section of the Acoustic model to map the spoken words to its respective audio file. This file plays a very important part in the final accuracy of the word detection as its 1:1-word mapping with each respective audio training file is crucial to perfecting the Acoustic model needed for accurate phone detection [20,21].

Tag	Word	Tag	Audio File Name
<s>	আট	</s>	aat
<s>	আছ্	</s>	acho
<s>	আকাশ	</s>	akash
<s>	আলো	</s>	alo
<s>	আমাদের	</s>	amader
<s>	আমার	</s>	amar
<s>	আমি	</s>	ami
<s>	আমরা	</s>	amra
<s>	বাতাস	</s>	batash
<s>	ভাই	</s>	bhai
<s>	ভাত	</s>	bhat

Figure 4.3: Transcription File for Training

4.1.5 Training and Testing

For training we wrote some commands to create language model, dictionary, phone set, file id and transcription file.. After running the whole system a file named other.html created, with that file we fixed the error. After training we create files for testing which are test.fileids and test.transcription. In test.fileids the directory for the test audio is attached. Similarly, in test.transcription the audio file and its actual word in Bengali is attached. After running the test code, it calculates the accuracy by matching the actual word in train.transcription with the test audio in test.transcription. Here figure 4.4 shows the test file IDs file and Figure 4.5 shows the transcription file for testing.

```
test/alobatash
test/amaderdhaka
test/amadernouka
test/amarghori
test/amibhat
test/amibhat2
test/amibujhi
test/amibujhi2
test/amikikori
test/amikikori2
test/amina
test/amina2
test/amra
test/amra2
test/amrakaj
test/amrakaj2
test/bikalshuru
test/bikalshuru2
test/choynodi
test/choynodi2
```

Figure 4.4: Audio File Directories From test.fileids

e	Word	Tag	Audio File Name
<s>	আলো বাতাস	</s>	alobatash
<s>	আমাদের ঢাকা	</s>	amaderdhaka
<s>	আমাদের নৌকা	</s>	amadernouka
<s>	আমার ঘড়ি	</s>	amarghori
<s>	আমি ভাত খাই	</s>	amibhat
<s>	আমি বুঝি	</s>	Amibujhi
<s>	আমি কি করি	</s>	Amikikori
<s>	আমরা কাজ করছি	</s>	amrakaj
<s>	বিকাল শুরু	</s>	Bikalshuru
<s>	এক দুই তিন	</s>	ekdui
<s>	কেমন আছ	</s>	kemonacho

Figure 4.5: Transcription File for Testing

4.2 Experimental Result and Analysis

4.2.1 Experimental Result

We tried to develop the system in such a way so that gradually it decreases its error and accuracy gets increased. For that purpose we focused on training the system with different speakers and with recordings recorded in different environment. While, testing for accuracy we also provided unknown speaker recordings just to see how accurate can the system for unknown speakers.

However, we described about the numbers of speakers, accuracy and other experimental issues through the table datum and bar chart datum. At first we started with very simple and few word recordings to implement the set up properly. In that case, after using few basic words such as first 10 Bengali integers we finally could get the output. Although we had to face difficulties such as trying 3 times in different computers after failing to set up the environment for the system, but we finally could develop the system successfully.

Moreover, out team then started working on recordings for more words since we got very poor accuracy at time. We even got the expected output after 5 times. So therefore, we the four members started to record new words for the training. We could record 61 words each

person. To get the better accuracy then we tried further and got much better accuracy than it was before.

Furthermore, we then recorded the same words 2 more times so that the system can recognize the words quickly with accuracy. Since we found better result in our 2nd experiment, we were expecting better this time as well. Since we had to do it quickly for some valid reasons we worked in different computers with the recordings. Unfortunately to work with that, since the Bengali typing software version was not similar, the words were generating different ASCII value which did not allow us to complete the training. After solving this issue, we completed the training and testing. Wonderfully it provided us further better result.

Finally in order to getting best possible result we increased the number of speakers with the number of Bengali words to train the system. Beside, we used speaker to record some specific Bengali sentences. After completing all of these attempt and experiments, we finally could increase the accuracy of the system which is was better than the first experiment. We also discovered if could increase enough words and sentences in the training phase, the system can even provide surprising accuracy rate.

Experiment	Training Phase		Testing Phase	
	No. of Speakers	Number of Utterance per word	Accuracy for Unknown Speakers (%)	Accuracy for Known Speakers (%)
01	1	1	7.69%	17.54%
02	3	3	30.77%	43.28%
03	4	4	44.90%	56.31%
04	6	14	54.08%	70.84%
05	8	20	59.01%	78.57%

Figure 4.6: Result Data

- Experiment 1: Dataset of 1 speaker containing 10 words of recording of 10 unique Bengali words.
- Experiment 2: Dataset of 3 speakers containing 10 unique words of recording with 3 utterances of each words.
- Experiment 3: Dataset of 4 speakers containing 61 unique words of recording with 4 utterances of each words.
- Experiment 4: Dataset of 6 speakers containing 101 unique words of recording with 14 utterances of each words.
- Experiment 5: Dataset of 8 speakers containing 101 unique words of recording with 20 utterances of each words.

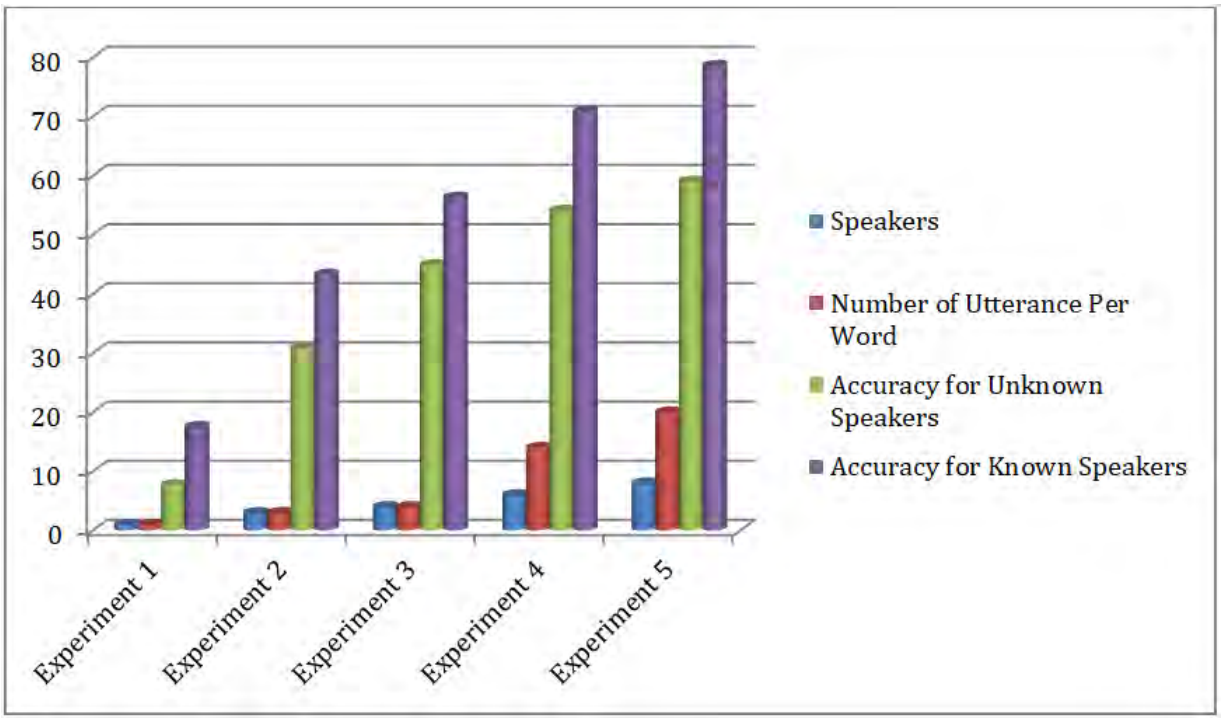

Figure 4.7: Accuracy Chart

To develop this complex system, we at first used thesis lab computers but we faced some technical difficulties. For that specific reason, we had to leave the lab and we started developing and experiment in our personal computer.

4.2.2 Analysis

CMUSphinx uses Word Error Rate (WER) to evaluate accuracy. WER is a very common way to calculate accuracy for speech recognition. Calculating WER is a way of measuring the number of errors that occurred during the translation of an audio signal. To calculate the error in our system, our first step was to record Bengali sentences by unknown and known speakers to the system. Then we put all the recordings in a different folder. After that we created a test fileids file, where we listed the directory and filename of each test recording. Then we also created a test transcription file where we listed all the test recordings' file names against the sentences that was recorded in the recordings. As a result, the system already knew what the correct output of each recording was. After everything was set, we started decoding the system with CMUSphinx. The decoding process used the acoustic model we trained and the language model we configured. When the recognition job was complete, the system generated the output from each of our test recordings. In the Figure 4.8, you can see the output that the system generated by speech recognition for each of the recordings.

Figure 4.8: Generated Output of each Test Recording

Then it compared between our test transcription file and its generated output. To calculate WER, the system first got the insertions, deletions, substitutions and the number of words in each recording. We can observe that in the below figure 4.9. Here, insertion is if the actual sentence was 3 words and if the generated output was 4 words. Then, then the value of insertion will be 1. Deletions is the exact opposite of insertions where the system calculates if the number of words in the generated output was less than the actual length of the sentence. In, substitution the system checks the number of wrong words in the output sentence. Then the system calculates the WER by the following equation,

$$WER = \frac{Insertions + Deletions + Substitutions}{TotalNumberofWords} \tag{4.1}$$

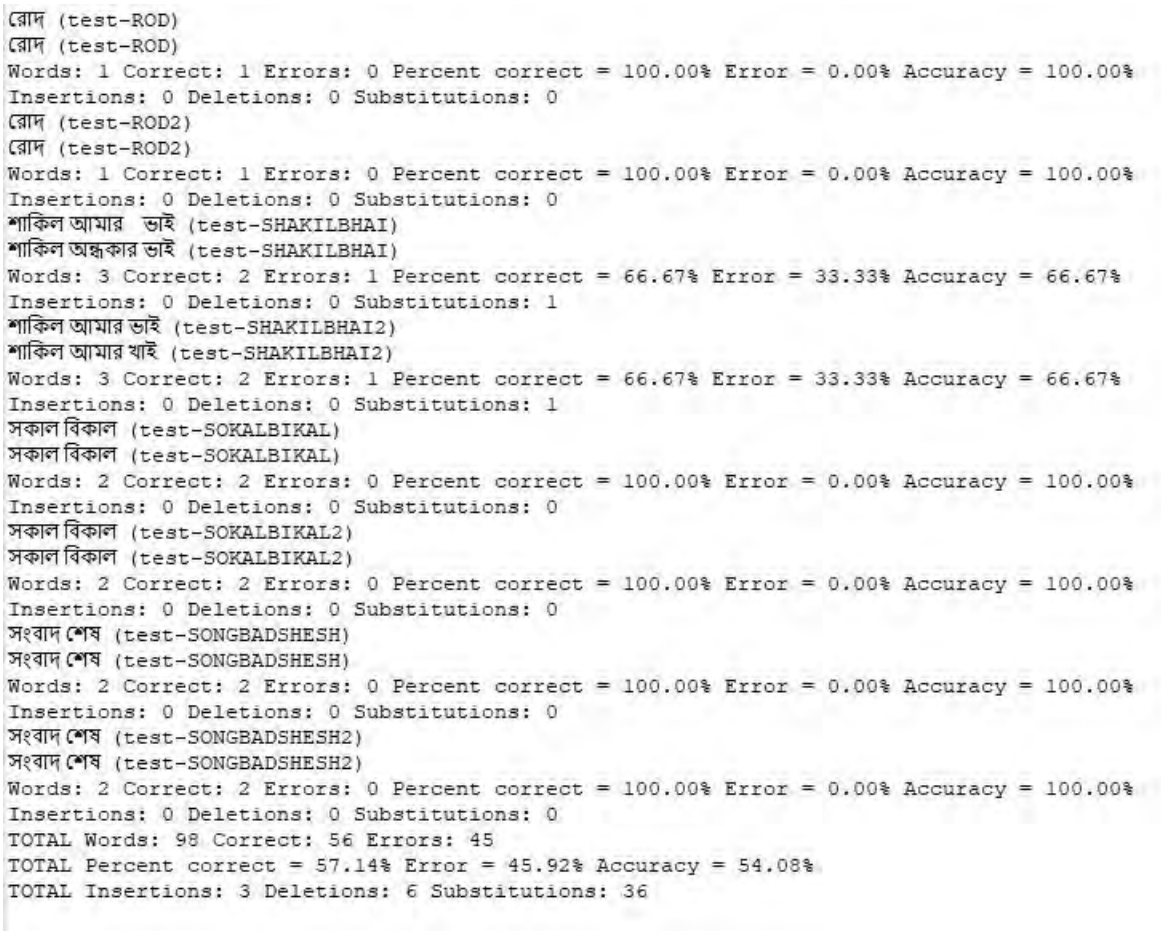

Figure 4.9: Accuracy of the Output

In our first attempt, we used only 10 Bengali words which were Bengali integers starting from 1 to 10. We used only 1 speaker recordings for the first experiment, we only got 7.69% of accuracy for unknown speakers and 17.54% of accuracy for known speakers, which let us know that our system works, even though the accuracy was poor. Then we added 2 more speakers and trained the system again with about 30 recordings and operated the testing. Completing the testing phase, we got better result which was about 30.77% for unknown speakers and 43.28% for known speakers. In this experiment we used 1 male speaker and 2 female speakers. However, since the accuracy rate was not enough and we only had 10 unique words, we decided to add more words to the dictionary. So, we decided to use 4 speakers

and added 61 words to the dictionary. Fortunately, this time our accuracy climbed up to 44.90% for unknown speakers and 56.31% for known speakers. Furthermore, when we added more speakers and recorded single word multiple times from each users we got better results and our accuracy increased from 44.90% to 54.08% for unknown speakers. After that, we enriched our dataset by adding 40 more words which makes our total words 101 and got the accuracy around 59.01%. For known speakers, we got 78.57% accuracy after testing. Our analysis shows that increasing the number of speakers and recordings helps to increase the accuracy but it has also limitation. Since, different people pronounce differently, it might not work for test speakers with different accent or pronunciation. But, according to Sphinx 4 documentation, the more data we could collect the better accuracy we could get from the result. Since it's really hard to collect such amount of datum from different individuals, the system may not perform 100%. But on the other hand, based on our collected data and training, after completing our testing phase we found that we can easily improve more if we can collect more recordings in the future.

4.2.3 Comparative Analysis of Different Models

In our thesis, some existing works on Speech to text conversion and speech recognition system are mentioned. Among them, there is a paper in which they convert speech by using SAPI[2] which is only worked on windows operating system as it is built with SAPI. Whereas, our proposed system runs on other operating systems besides windows such as Linux and ios. Moreover, our proposed system shows output in real time but in their system, first it takes input and stores it in a file then it shows the output.

There is a paper which recognizes Bangla real number automatically using CMU Sphinx4[12] implemented recognition system only for Bangla real number where our system we work for 101 bengali words including 10 bengali integers.

Our approcah for builting the proposed system is better than existing models as it is system independent. Furthermore, Our system can be easily implemented on mobile platform. Moreover it takes less time than others and runs in real time for continuous speech.

Chapter 5

Conclusion

5.1 Conclusion

To conclude, the paper we are presenting is a model of "voice to text conversion method" for Bengali language. We have already mentioned about that the system, which would be very helpful for people who are technically challenged or cannot type. It is also a game-changing tool for the illiterate people. It aims to help the deaf people by giving assistance to the academic sector. So the model is not only a model for mechanism but also effective and helpful for various purposes. This independent recognition system for continuous speech is built with an open source frame work of CMUSphinx which consists both male and female voices that are being recorded. With the help of word matching algorithm, the system recognizes the word and follows the steps to convert in process. To make the system more accurate, lots of improvement is required. Among of them, in our thesis, 4 speakers recorded each word 3 times and 2 speakers recorded each word 1 time and got the accuracy around 54.08%. The recognition results produced by our system showed to be satisfactory.

5.2 Future Work

We have a plan with this system to enhance in a different ways. For instance, we have an idea to add more about the rule based phonetic dictionary which will help in further variations. We want to add up more speakers that the accuracy rate might increase up to 100%. We would be trying to enhance the parameterization of the acoustic model. The main plan or future enhancement of our thesis is to build a Bangla Dictation Notepad Software which can be easily handled and all type of utterance and pronunciation can be modified. For further enhancement, there will be a Bangla dictionary for the mobile system with enriching new words. We want an integration of the system with Interactive voice response with the increase of its capability to recognize speech more accurately.

5.3 Limitations

The proposed model has some limitations. The first and foremost one is the accuracy of the system.

While training and testing. we trained the system with 10 words at the first attempt. At that attempt the system accuracy was only 7.69% which was for only 1 speaker. Then in final attempt, after increasing the number of speaker to 8 and the words to 101 the system accuracy climbed upto 59.01%. The accuracy could be increased more than we got in our system by adding more speakers and number of words. We believe that if the number of speakers are increased, the accuracy could be increased easily which we noticed from the first attempt to final one.

Moreover, we worked with a framework(CMUSphinx), in which Bengali speech to text conversion was not done in a vast perspective. Because of that we could not get any help or use previous dataset which could have helped us to decrease our time on training the system. We created the whole dataset newly from the scratch with unique speakers and words. We also believe that if there was previous dataset that we could use, then we could easily increase the accuracy as well.

Bibliography

1. Chowdhury, Shammur. (2010). Implementation of speech recognition system for Bangla. 10.13140/RG.2.1.4658.4488.

2. S. Sultana, M. A. H. Akhand, P. K. Das, M. M. H. Rahman, "Bangla Speech-to-Text Conversion using SAPI," In Computer and Communication Engineering (ICCCE), International Conference. Kuala Lumpur, 3-5 July 2012. Kuala Lumpur: IEEE. P.385-3, 2012.

3. K.M. Shivakumar, K. G. Aravind, T. V. Anoop, G. Deepa "Kannada Speech to Text Conversion Using CMU Sphinx," International Conference on Inventive Computation Technologies (ICICT), Vol 3, 2016.

4. Nasib, Abdullah Kabir, Humayun Ahmed, Ruhan Uddin, Jia. (2018). A Real Time Speech to Text Conversion Technique for Bengali Language. 1-4. 10.1109/IC4ME2.2018 .8465680.

5. Dana Dannells, "Disfluency detection in a dialogue system".

6. Mishra, N., Shrawankar, U., Thakare, V. M. (2013). An Overview of Hindi Speech Recognition. arXiv preprint arXiv:1305.2847.

7. Satori, H., Harti, M., Chenfour, N. (2007). Introduction to Arabic speech recognition using CMU Sphinx system.

8. Liu, Y., Shriberg, E., Stolcke, A. (2003, September). Automatic disfluency identification in conversational speech using multiple knowledge sources. In INTERSPEECH.

9. Honal, M., Schultz, T. (2003, April). Correction of disfluencies in spontaneous speech using a noisy-channel approach. In INTERSPEECH.

10. Tsvetkov, Y., Sheikh, Z., Metze, F. (2013, May). Identification and modeling of word fragments in spontaneous speech. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7624-7628). IEEE.

11. Vadwala, Ms.Ayushi Suthar, Ms.Krina Karmakar, Ms.Yesha Thakkar, Nirali. (2017). Survey paper on Different Speech Recognition Algorithm: Challenges and Techniques.

12. Nahid, Md Mahadi Islam, Md Islam, Md Saiful. (2016). A noble approach for recognizing Bangla real number automatically using CMU Sphinx4. 844-849. 10.1109/ICIEV .2016.7760121.

13. El Amrani, Mohamed Yassine Rahman, M. M. Ridza Wahiddin, Mohamed Shah, Asadullah. (2016). Towards Using CMU Sphinx Tools for the Holy Quran Recitation Verification. International Journal on Islamic Applications in Computer Science And Technology. 4.

14. Ghadage, Y. and Shelke, S. (2016). Speech to text conversion for multilingual languages - IEEE Conference Publication.

15. Songfang Huang and S. Renals, "Hierarchical Bayesian Language Models for Conversational Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 8, pp. 1941-1954, 2010. Available: 10.1109/tasl.2010.2040782.

16. B. TombaloĞlu and H. Erdem, "Survey on Phoneme Recognition using Support Vector Machine", Issuu, 2017. [Accessed: 15- May- 2017].

17. A. Bapat and L. Nagalkar, "Phonetic Speech Analysis for Speech to Text Conversion - IEEE Conference Publication", Ieeexplore.ieee.org, 2008 [Accessed: 06- Mar- 2009].

18. R. Shadiev et al., "Applying Speech-to-Text Recognition and Computer-Aided Translation for Supporting Multi-lingual Communications in Cross-Cultural Learning Project", 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 2017 [Accessed: 08- Aug- 2017].

19. Shmyrev, N. (2019). Overview of the CMUSphinx toolkit. [online] CMUSphinx Open Source Speech Recognition. Available at: https://cmusphinx. github.io/wiki/tutorialoverview/ [Accessed 3 Aug. 2019].

20. 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].

21. 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].

22. 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.

23. Md. R. Mijanur, Md. K. Farukuzzaman, A. M. Mohammad, "Speech Recognition Front-end for Segmentation and Clustering Continuous Bangla Speech," DIU Journal of Science and Technology, vol 5, Issue 1, January 2010.

24. Sultana, Rumia Palit, 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.

25. Oliveira, Rafael Batista, Pedro Neto, Nelson Klautau, Aldebaro. (2012). Baseline Acoustic Models for Brazilian Portuguese Using CMU Sphinx Tools. 7243. 375-380. 10.1007/978-3-642-28885-242.

26. Phull, Disha Bharadwaja Kumar, G. (2016). Investigation of Indian English speech recognition using CMU sphinx. 11. 4167-4174.

27. 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.

28. Wang, W., Liao, Y. and Chen, S. (2002). RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion. Speech Communication, 36(3-4), pp.247-265.

29. Maegawa, S. (2012). SPEECH RECOGNITION METHOD, SPEECH RECOGNITION SYSTEM AND SERVER THEREOF. The Journal of the Acoustical Society of America, 131(6), p.4868.

30. Sean R. Eddy, "Multiple alignment using hidden Markov models," Dept. of Genetics, Washington University School of Medicine, 1995.

31. G.F. David Jr, "The Viterbi Algorithm," Proceedings of the IEEE, vol. 61, No. 3, March 1973.

32. W. Adam, "Digital Audio Workstations and Analog to Digital Conversion," Department of Computer Science (CIS) University ofWisconsin-Platteville.

33. N. Francis, "Intonational equivalence: an experimental evaluation of pitch scales," Department of Linguistics, University of Cambridge, UK

34. Mohamed, A. R., Dahl, G. E., and Hinton, G., "Acoustic Modelling using Deep Belief Networks",submitted to IEEE TRANS. On audio, speech, and language processing, 2010.

35. Sorensen, J., and Allauzen, C., "Unary data structures for Language Models", INTERSPEECH 2011.

36. Kain, A., Hosom, J. P., Ferguson, S. H., Bush, B., "Creating a speech corpus with semi-spontaneous, parallel conversational and clear speech",Tech Report: CSLU-11-003, August 2011.