Speech to Text Recognition

Al Noman Takrim	**Id: 011-111-072**
Md. Kamruzzaman	**Id: 011-133-090**
Ananna Zoha	**Id: 011-133-079**
Sabbir Ahmmad	**Id: 011-133-095**

A thesis in the Department of Computer Science and Engineering presented

In partial fulfillment of the requirements for the Degree of

Bachelor Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

July, 2018

Declaration

We declare that this thesis titled, Speech to Text Recognition and the work presented in it are my own. We confirm that:

- This work was done wholly or mainly while in candidature for a [BSc]* degree at United International University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

Al Noman Takrim
ID: 011-111-072
Computer Science and Engineering

Md.Kamruzzaman
ID: 011-133-090
Computer Science and Engineering

Ananna Zoha
ID: 011-133-079
Computer Science and Engineering

Sabbir Ahmmad
ID: 011-133-095
Computer Science and Engineering

Certificate

I do hereby declare that the research works embodied in this thesis entitled "**Speech to Text Recognition**" is the outcome of an original work carried out by Al Noman Takrim, ID: 011-111-072, Md. Kamruzzaman, ID: 011-133-090, Ananna Zoha, ID: 011-133-079, Sabbir Ahmmad, ID: 011-133-095 under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of BSc in Computer Science and Engineering.

Dr. Mohammad Nurul Huda
Professor and Coordinator of MSCSE Program
United International University

Abstract

In terms of communication between human races, the Speech is the most appropriate mode. When this communication occurs with human and computer this is called human computer interface. To interact with computer, speech could be the most efficient medium. This paper gives pure knowledge of how speech could be converted into text. This paper will be helpful to know more about evolution of speech to text recognition. How todays modern technologies has been cope up with voice recognition system, in this paper those terms are being explained very carefully. This paper also describes how a user interface could be build up for voice recognition. There is a lot of existing speech recognition software. Microsoft, Google, Nuance Communications, IBM, they all have speech recognition software. In terms of modern technology speech recognition system is way too compatible. Someone should speak, and there is a program where there is a list of grammar. When speaker speaks, the hardware receives the voice and processes it with the program. Based on training and grammar, the system identifies the voice precisely and converts them into recognized text.

Acknowledgement

This thesis paper was supervised by Dr. Mohammad Nurul Huda, Professor and Coordinator of MSCSE program, United International University. We thank our beloved professor for helping us to complete such a great work. I, Al Noman Takrim, ID: 011-111-072 thank my team members Md. Kamruzzaman, ID: 011-133-090, Ananna Zoha, ID: 011-133-079, Sabbir Ahmmad, ID: 011-133-095. Without their hard work this might not be possible to complete such a work.

Table of Contents

LIST OF TABLE

LIST OF FIGURES

Chapter1

Introduction

Speech Recognition is a modern system where speech can be recognized automatically by the computer and convert them into text. By converting speech into text one can control digital devices by speaking instead of writing or using keyboards. This reduces a lot of time and work. It makes human life easier.

Did you ever talked to a computer? Where the computer really recognized your voice and converted them into text. Suppose not. The main focus of speech recognition system is to process spoken input and translate them into text that a program understands well. The software system interprets the outcome of the result of the recognition as a command. So you can consider it as a command and control application. For an example say “What time is it?” The program will recognize what you said. Then the program will convert them into text and command the program to execute. Then the program will reply the time for you. But how does it actually work? How a recognition speech understands that there is a speech. For that the system has to identify the utterances.

Utterance is the source of speech between two periods of silence. This utterance is sent to the speech engine. Then the system needs to know the pronunciation of the word. If the system is not trained by the pronunciation, then it won’t be able to identify what a speaker said. This is the evolution of modern science.

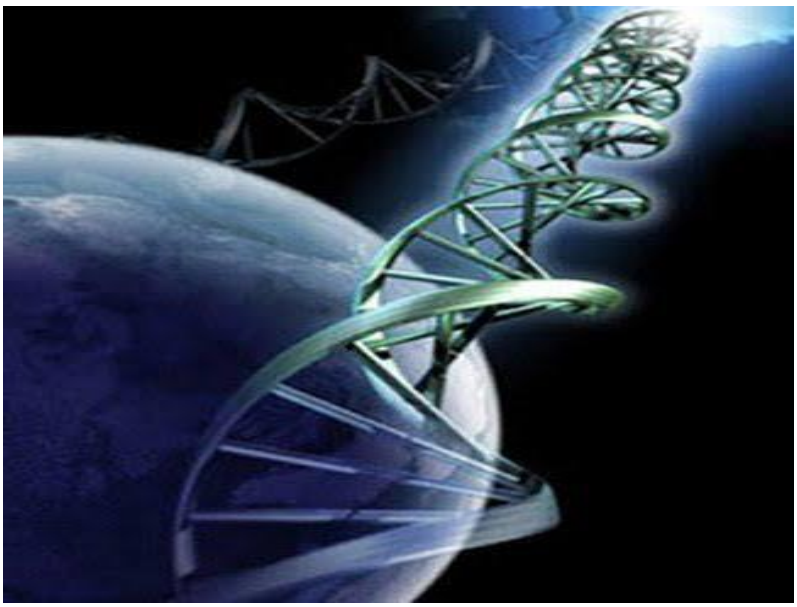

Figure 1: Evolution of Science

Chapter 2

Background and Literature Review

2.1 Definition

Speech Recognition is the ability of a machine or a program to identify the words in spoken language and convert them into text. Speech to text system is an inter-disciplinary sub-field of computational linguistics. This system is also called as Automatic Speech Recognition or ASR. In short this is called STT as well. Speech recognition system develops methods and technologies that enable the recognition and conversion of spoken language into text.

2.2 View

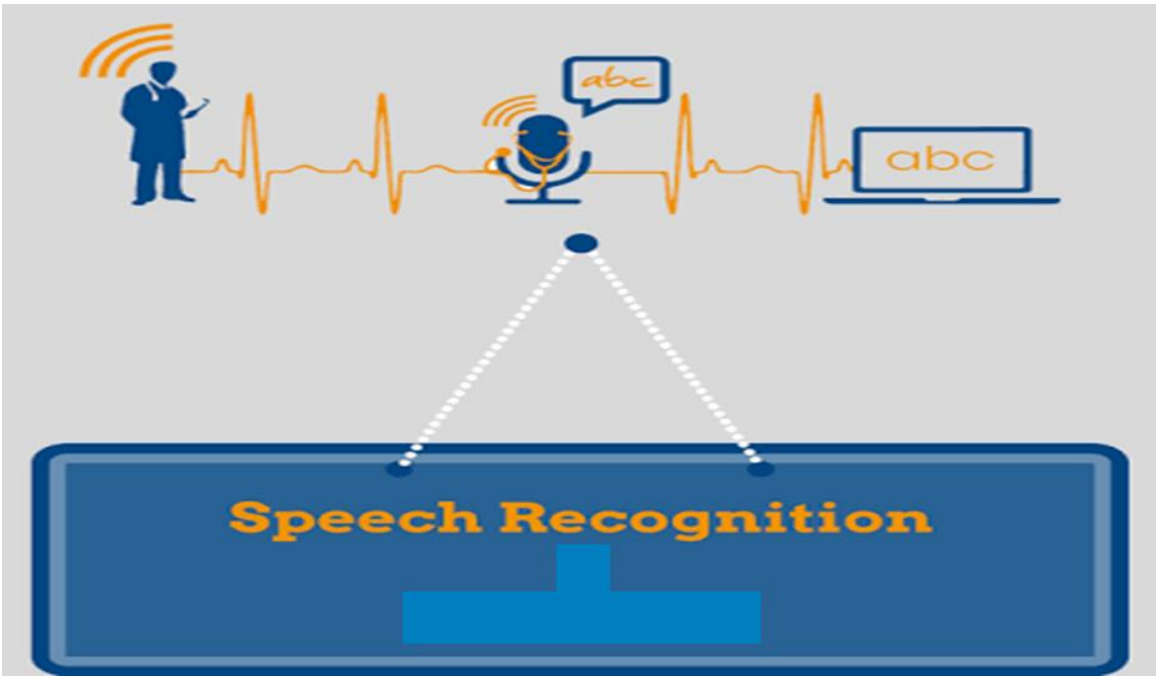

Figure 2.2.1: Speech to Text Recognition

2.3 Requirements

At first we have to train our system. This is called “Training”. This process is also called “Enrollment”. There should be individual speaker to speak. The system will be needed a computer or a program to receive the speech precisely. The computer should have the ability to perform the action. Sometimes we can use vocabulary too.

2.4 Early Work

In 1950, the speech to text recognition system was limited to single-speaker system consists of around ten words. The vocabulary was very poor that time. In 1952, the system was designed for single-speaker digit recognition. In 1994, Nuance Technology was founded as a spinoff of SRI International’s Speech technology and Research (STAR) Laboratory to commercialize a speaker independent system. Nuance launched their first speech application in 1996. After that the speech recognition system was enriched day by day.

2.5 Modern Systems

Figure 2.5.1: Modern Systems

In early 2000s, the speech recognition system was dominated by traditional approaches. Researchers worked with Hidden Markov Model combined with Neural Networks. After that it was taken by a deep learning method called LSTM. Around 2007, Connectionist Temporal Classification outperformed the traditional speech recognition system. In 2015, Google made a system which jumped of 49% through CTC-trained LSTM.

2.6 Applications

There are so many applications of speech to text recognition in the present time. Most of the applications are being used very successfully. Speech to Text Recognition system helped us in many ways. You can do anything only by speaking nowadays. Speech Recognition system has numerous fields. Dictation, Text-To-Speech, Command Input, Translation these are the sub-field of speech recognition.

Figure 2.6.1: Home Automation System, Mobile Phone, IVR

1. People with disabilities.
2. Education and daily Life.
3. Telephony and other domains.
4. Training air traffic controllers.
5. Military.
6. Health care.
7. Automatic translation.
8. Pronunciation.
9. Home automatic system.
10. Virtual assistance.
11. Interactive voice response.
12. Automatic subtitling with speech recognition.

2.7 Usefulness

By Speech to text recognition our work processes become more efficient. When it comes to document processing the time is shorter. Various kinds of work can be generated up to 3 times faster with speech to text recognition. It saves a great deal of labor. When the system identifies any error and the error is corrected, the system learns that automatically. As a result the recognition rate is more precise than ever. Speech recognition system allows dictations as well. In online business like shopping, restaurant, organizations like Amazon, Ali Baba, FlipKart have different department for answering customers.

1. Efficient work processes.
2. Time saver.
3. Auto correction.
4. Dictations.
5. Cost effective.
6. Less manpower.
7. Easy to evaluate.
8. Easy to use.
9. User friendly.
10. Speaker independent.

Figure 2.7.1: Reduced time and hard work

2.8 Limitations

When building our system we faced several difficulties. It was very challenging for our team to overcome those difficulties. Moreover, the system can't translate from one language to another. The speech recognition system needs training and enrollment. So without training the system won't be able to work. The system won't work on any platform other than windows like mac, ubuntu etc. Speech recognition relies on two components. Number one is the language model. But it is very expensive and time consuming to establish such a language model. That is why the language model of speech recognition is specific. The system can't work with random language. Which language is used to train the system, will depend on that system. Number two is the acoustic model. Sometimes the same language can be confusing. For example, US English, UK English has different regional accents. These models take lots of lots of training. There are some other difficulties too.

1. Set-up and training was difficult.
2. It was very time consuming.
3. We had limited vocabulary.
4. Difficulties making the user interface design.
5. Difficulties during HTK simulations.
6. Difficulties receiving the voice.
7. Difficulties in noise reduction.
8. Speech engine system is platform dependent.

CHAPTER 3

Proposed System

3.1 Hidden Markov Model

Programmed persistent discourse acknowledgment (CSR) has numerous potential applications including charge and control, correspondence, interpretation of recorded discourse, looking sound archives and intelligent talked discoursed. The center of all discourse acknowledgment frameworks comprises of a set of factual models speaking to the different hints of the dialect to be perceived. Since discourse has worldly structure and can be encoded as a succession of phantom vectors crossing the sound recurrence go, the concealed Markov demonstrate (HMM) gives a characteristic structure to developing such models.

Hidden Markov Model is the core of all modern speech recognition system. This model is a statistical Markov Model. The HMM can be represented as the simplest dynamic Bayesian Network. When you use Markov model, it is only visible to the observer but when you use HMM the states are not visible to the observer. It provides simple and effective framework for modeling.

The foundation of modern Hidden Markov Model or the basic framework of HMM has not been changed significantly I last decade or more. The further modeling technique has been developed within this framework.

3.2 Architecture

The chief segments of an extensive vocabulary constant discourse recognizer are outlined in Figure 3.2.1. The info sound waveform from a mouthpiece is changed over into a succession acoustic vectors in highlight extraction. The decoder at that point discover the arrangement of words w1:L = w1,...,wL which is well on the way to have created Y , i.e. the decoder attempts to discover

In any case, since P(w|Y) is hard to,1 Bayes' Rule is used to change into the proportionate issue of finding**:**

The is by an acoustic model and the earlier P(w) is dictated by a dialect display. The unit of sound There are a few frameworks that depend on discriminative models where P(w|Y) is demonstrated straight forwardly, as opposed to utilizing generative models, for example, HMMs where the is displayed, p(Y |w). by and by, the isn't standardized and the dialect show is regularly scaled by an observationally decided consistent and a word inclusion punishment is included i.e., in the log area the is computed as log where α is ordinarily in the range 8– 20 and β is regularly in the range 0-20.
HMM works with technique. The system receives the speech via microphone. After receiving the speech, the first task is to extract the feature from the speech. Feature is like pronunciation, accent etc. After that the feature is given to the decoder. Decoder decodes those feature based on three parameters, Acoustic model, Pronunciation Dictionary, Language model. Then after decoding based on those parameters the recognized words are shown into output sequence.

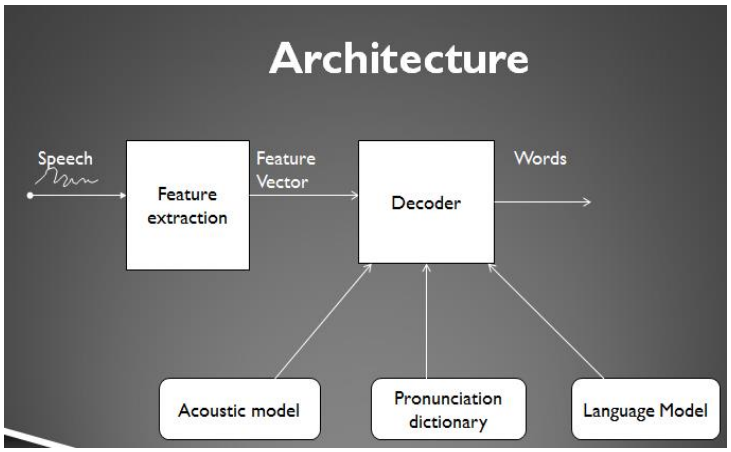

Figure 3.2.1: Architecture of a HMM based Recognizer

3.3 Feature Extraction

The feature extracted from those data seeks to provide a compact representation of speech waveform. The waveform created from those feature extraction should minimize the loss of information. To identify the words precisely and make the perfect recognition this feature extraction is very important. For example, if diagonal covariance Gaussian distributions are used for the state-output distributions then the features should be designed to be Gaussian and uncorrelated.

The component extraction arranged tries to give a reduced portrayal of the discourse waveform. This shape ought to limit the loss of data that separates amongst words, and give a decent coordinate with the distributional suppositions made by the acoustic models. For instance, if slanting covariance Gaussian conveyances are utilized for the state-yield dispersions then the highlights ought to be outlined to be Gaussian and

uncorrelated. Highlight vectors are commonly figured each 10 ms utilizing a covering examination window of around 25 ms. One of the least complex and most generally utilized encoding plans depends on mel-recurrence cepstral coefficients (MFCCs) . These are produced by applying a truncated discrete cosine change (DCT) to a log ghastly gauge registered by smoothing a FFT with around 20 recurrence canisters circulated non-directly over the discourse range. The nonlinear recurrence scale utilized is known as a Mel scale and it approximates the reaction of the human ear. The DCT is connected all together to smooth the gauge and roughly decor relates the highlight components. PLP figures direct expectation coefficients from a non-directly compacted control range and afterward changes the direct expectation coefficients to cepstral coefficients. Practically speaking, PLP can give little changes over MFCCs, particularly in loud situations and consequently it is the favored encoding for some frameworks first request (delta) and second-arrange (delta– delta) relapse coefficients are frequently annexed in a heuristic endeavor to make up for the restrictive autonomy presumption made by the HMM-based acoustic models.

Feature vectors are counted in every 10ms using overlapping window of 25ms. The most widely used encoding scheme depends on mel-frequency cepstral coefficients (MFCCs). These are built by truncated discrete cosine transformation (DCT) to a log spectral. Another scale used as a non-linear frequency scale is called as mel scale. When the cosine transformation is done number one element presents average of log-energy of frequency bins.

3.4 System Design

Our system design describes the whole work of our papers. The design shows that at first speaker sends the speech signal to the microphone. The device receives the signal and takes it to the pre-processing unit. After pre-processing the feature vectors or the signal, the signal is given to mel-frequency cepstral coefficients. This MFCC analyze the feature vectors. Then some of it goes to Code Book Generation and some to the direct vector quantization. The feature model given to CBG goes to the code book. Then again a model created by code book is given to vector quantization. Then it directly goes to HMM Recognition. During the vector quantization some vector is trained by HMM. HMM makes training data to train the model. Then the model goes to HMM Recognition. The whole system identifies the speech and converts them into text very precisely.

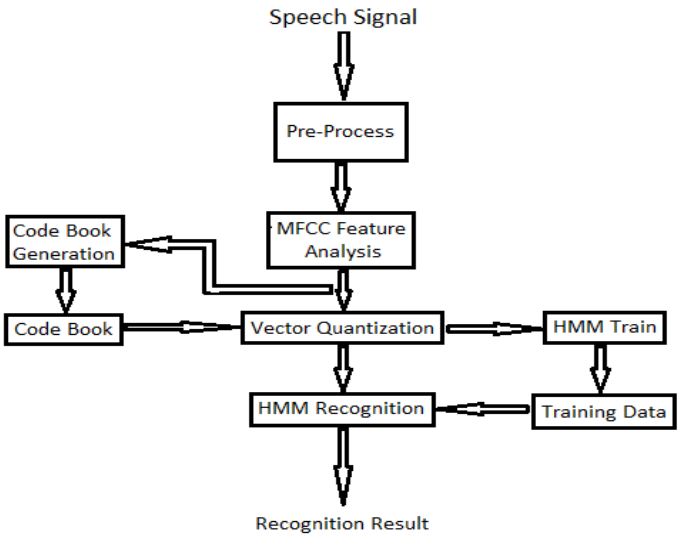

Figure 3.3.1: System Design

3.5 Language Models

The earlier likelihood of a word arrangement w = w1,...,wKrequired in is given by

P(w) = K k=1 P(wk|wk−1,...,w1) (3.7.1)

For expansive, the molding word history is generally truncated to N−1 words to shape a N-gram dialect display.

Where N is regularly in the range 2–4. Dialect models are regularly evaluated as far as their perplexity H, which is characterized.

Where the estimation is utilized for models with a limited length word succession. N-gram Language Models. The N-gram probabilities are assessed from preparing writings by tallying to shape most extreme parameter gauges. For instance, let C(wk−2wk−1wk) speak to the number of the three words and comparatively for C(wk−2wk−1), at that point.

The significant issue with this basic ML estimation plot is information sparsely. This can be moderated by a blend of marking down and backing-off. For instance, utilizing alleged is a check edged is a α is a

In this way, at the point when check some edge utilized. At the point when the check is little a similar ML evaluate is utilized yet. The marked down

which are by a of the relating bigram. This thought can be connected recursively to appraise any meager N-gram regarding an arrangement of and (N − 1) grams. The marking down depends on the Turing-Good where nr is the quantity of N-grams that happen precisely r times in the. There are numerous varieties on this approach. For

instance, when preparing information is extremely scanty, Kneser–Ney smoothing is especially successful. An elective way to deal with vigorous dialect demonstrate estimation is to utilize class-based models in which for each word wk there is a comparing class ck .

Again look over the vocabulary, testing each word to check whether moving it to some different class would build the probability.

By and by it is discovered that for sensibly measured preparing sets,8 an compelling dialect show for extensive vocabulary applications comprises of a smoothed word-based 3 or 4-gram added with a class-based trigram.

3.6 UI Design

We made a system for the users in a very easy way. Our interface is very user friendly. Anyone can operate our User Interface. We used Sphinx to build this user interface. You can see that we simply kept three buttons in our system software. There are three buttons, Start, Pause, Resume. If you want to say something just grab the microphone, press the start button and start talking. When you are done talking your speech will be converted automatically into text and will be shown in the white box.

If you want to give a pause while talking, you just press the pause button the whole system will be paused. You can start your speech from where you left off. So you can see that this user interface is easy to operate. Anyone can use this software and have the fun to see something magical.

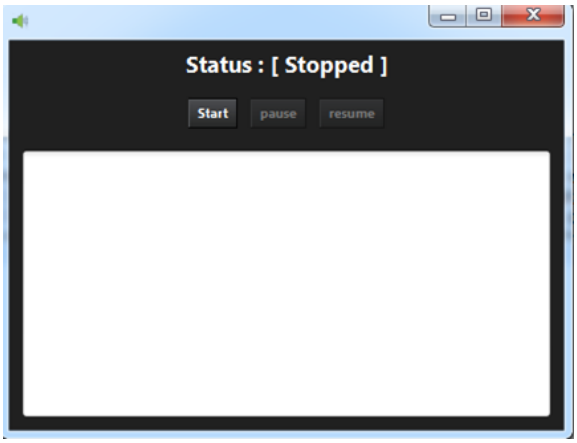

Figure 3.4.1: User Interface- i

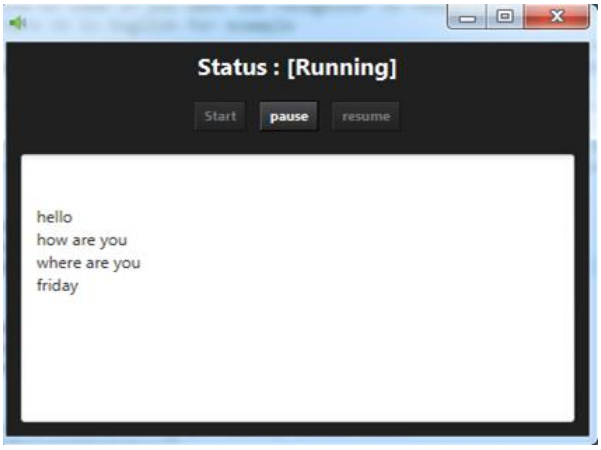

Figure 3.4.2: User Interface- ii

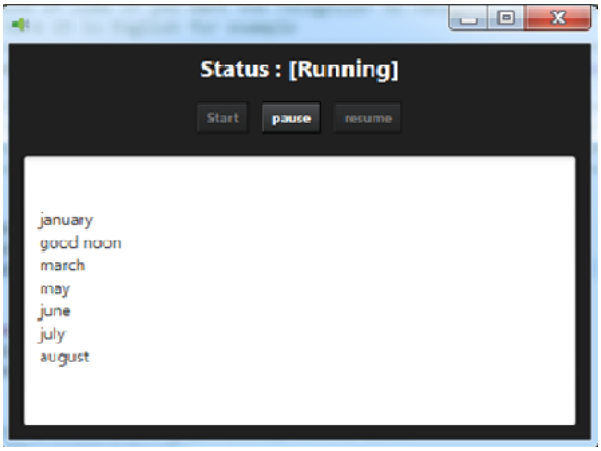

Figure 3.4.3: User Interface- iii

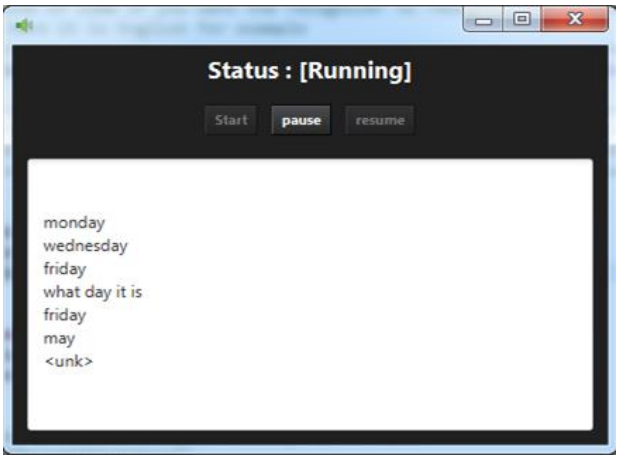

Figure 3.4.4: User Interface- iv

3.7 Experiments: Corpus

At first while experimenting we just used random speaker. Anyone could speak. Those words are not present in our vocabulary or dictionary. We just wanted to see that without dictionary how well our system captures the speech and convert them into text. But unfortunately we noticed that because of several limitations the system is not doing great to identify the speech. There is so much noise around us. Our noise reduction system is not too good. Moreover, we didn't train our system with the dictionary by which our system will be able to match the speech with the dictionary.

3.8 Experiments: Dictionary

Initially we used corpus to experiment our system. But because of various kinds of difficulties like noise, receiving speech etc. the system was unable to identify the words precisely. So we used a set of vocabulary to help the system to identify the words more precisely. The vocabulary list is given below:

1. where are you
2. how are you
3. I am going to school
4. they are playing football
5. paragraph
6. mathematician
7. zoology
8. hello
9. good
10. mango
11. which day it is
12. how are you
13. who are you
14. what's your name
15. I eat rice
16. I go to school
17. my name is X
18. this is a table
19. I go to school
20. we play football
21. he is playing
22. the baby is crying
23. I want to go home
24. this is an elephant
25. United International University
26. Professor

3.9 Experiments: Abbreviation

We tried all possible vocabularies. But our system was not able to identify the short forms instead of adding them into dictionary. Some of those examples are given below:

1. UIU
2. CSE
3. EEE
4. CTE
5. ECE
6. CPU
7. DIY
8. EOM
9. FAQ
10. HTF
11. IDK
12. IOT
13. ITT
14. N/A
15. OMG
16. POV
17. TBA
18. TIA
19. TQ
20. WTT
21. YAM
22. WRT

These kinds of short forms were not identified by our system . But we are working on it to work properly. We will make build something that will be able to identify and convert anything in the world.

CHAPTER 4

Results

The above experiments have described a number of points about large vocabulary for speech recognition. In table 4.1 the results of our experiments have been discussed.

ID	Experiment	Number of unit	Accuracy [%]
1.	Corpus	200	37.39
2.	Dictionary	200	68
3.	Abbreviation	200	9.41

Table 4.1: Recognition Accuracy Rate

Above table describes the recognition rate of our system. Based on the table we can see that when we worked with the corpus we used 200 set of words. Our system's successful recognition is around 38%. But when we trained our model with 200 vocabularies our system recognized 68% successfully. On the other hand when we worked with abbreviation we got only 9.41% success rate that is quiet poor recognition rate. So we can say that our system works better with trained model.

CHAPTER 5

Summary and Conclusion

5.1 Summary

The above analyses have represented various focuses about expansive vocabulary consistent. Regarding the basic discourse, the same wide arrangement of strategies can be utilized for various dialects, for this situation English, Arabic and Mandarin. Be that as it may, to empower the innovation to for various dialects, adjustments telephone units and word reference are required. For instance in Mandarin, division must be considered, and in addition of tonal highlights. The frameworks portrayed above have all been founded on HLDA / correlation. This, , for example, worldwide STC, are standard methodologies utilized by a large number of the exploration bunches creating huge vocabulary frameworks. Moreover, all the been discriminative, MPE, preparing. These discriminative methodologies are additionally getting to be standard since run of the of around 10% relative contrasted with ML can be gotten.

Despite the fact that no given, every one of the frameworks depicted make utilization of a blend of highlight standardization and direct adjustment plans. For substantial vocabulary discourse acknowledgment errands test information may involve a scope of talking styles and emphasizes, these procedures are basic. As the outcomes have picks up can be by consolidating together. For instance in Arabic, emic and phonetic together, additions of 6% relative over the gotten. Every one of the portrayed here were worked at CUED. Curiously increases can cross-site framework blend. Despite the fact that BN and BC translation is a decent errand to show numerous of the plans in this survey, rather unique undertakings would be expected to the heartiness systems depicted and space anticipated of these. Some plans have been connected to, in any case, the picks up have been little since the levels in this sort of interpretation are ordinarily low.

The tests portrayed have likewise maintained a strategic distance from issues of computational effectiveness. For commonsense frameworks this is an imperative thought, in any case, past the extent of this audit.

5.2 Conclusion

Speech to Text Recognition is such a system which has a great impact on modern times. Nowadays, you can do anything through which helps you to save money, time and effort. Our proposed and build system has the potential to serve you in great ways. HMM based relies on a set of assumptions. We can say that speech signals can be represented as a set of feature vectors. And can be trained by HMM. By HMM feature vectors can be distributed and modeled exactly. By those feature vectors we get training data and those training data are enough to recognize the speech and convert them into text.

There are some other speech recognition. But it varies because of availability of training data or adaptation data. Such a system can be threatened by run time computation, system complexity and target application. To build a successful large vocabulary transcription it's always a challenge. So the more we can train our model, the more we will get accuracy of recognition. Without training the system won't be able to recognize the speech perfectly.

This audit has surveyed the center design of sketched out the real zones of refinement fused into cutting edge frameworks including those intended to meet the requesting errand of extensive vocabulary Well acknowledgment frameworks depend with respect to various presumptions: that discourse signs can be spoken to by a succession of frightfully inferred include vectors; that can be utilizing a HMM; that the can be displayed; and that conditions are coordinated. Practically speaking, must be casual to some degree and it is the degree to which the subsequent approximations can be limited which decides execution.

Beginning from a base of basic slanting covariance ceaseless thickness HMMs, a scope of refinements have been depicted counting dissemination and covariance displaying, discriminative parameter estimation, and calculations for adjustment and clamor pay.

These mean to lessen the fundamental HMM structure and consequently enhance execution. Much of the time have been portrayed among which there was frequently. For the most part, an exchange off between elements, for example, the accessibility information, run-time calculation, and target application. The outcome of this is building a fruitful isn't simply about finding rich answers for. It is likewise a testing issue. As represented by the last introduction of multi-pass designs and real illustration arrangements, the most require complex arrangements. Well based acknowledgment is and as prove by the organization, execution has just achieved a level which can bolster. Advance is ceaseless and on falling. For illustration, on the translation of conversational phone discourse were around half in 1995. Today, with the, furthermore, the depicted in this audit, blunder rates are currently well beneath 20%. By the by. As appeared by the illustration given for interpretation of, comes about are amazing yet at concerning human capacity. Besides, even the best frameworks are powerless against unconstrained talking styles, non-local or and high surrounding commotion. Luckily, in this audit which

presently can't seem to be completely investigated and more still to be considered. This is certainly valid, no great option to the HMM has been found yet. Meanwhile, of this survey trust that the execution asymptote for discourse acknowledgment is still some way away.

5.3 Future Development

In future we will try to make our system more precise. We will be able to identify anything that a speaker says. Our system works only in English. But now we will focus on how to make a system where speaker can speak in Bengali and our system should identify the word and convert them into Bengali text.

6. References

[1] Jingdong Chen, Member, Yiteng (Arden) Huang, Qi Li, Kuldip K. Paliwal, "Recognition of Noisy Speech using Dynamic Spectral Subband Centroids" IEEE SSIGNAL PROCESSING LETTERS, Vol. 11, Number 2, February 2004.

[2] Hakan Erdogan, Ruhi Sarikaya, Yuqing Gao, "Using semantic analysis to improve speech recognition performance" Computer Speech and Language, ELSEVIER 2005.

[3] Chadawan Ittichaichareon, Patiyuth Pramkeaw, "Improving MFCC-based Speech Classification with FIR Filter" International Conference on Computer Graphics, Simulation and Modelling (ICGSM"2012) July 28-29, 2012 Pattaya(Thailand).

[4] Bhupinder Singh, Neha Kapur, Puneet Kaur "Sppech Recognition with Hidden Markov Model:A Review" International Journal of Advanced Research in Computer and Software Engineering, Vol. 2, Issue 3, March 2012.

[5] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition using MFCC and DTW" International Journal of Advance Research in Electrical, Electronics and Instrumentation Engineering, Vol.2, Issue 8, August 2013.

[6] Ibrahim Patel, Dr. Y. Srinivas Rao, "Speech Recognition using HMM with MFCC-an analysis using Frequency Spectral Decomposition Technique" Signal and Image Processing:An International Journal(SIPIJ), Vol.1, Number.2, December 2010.

[7] Om Prakash Prabhakar, Navneet Kumar Sahu,"A Survey on Voice Command Recognition Technique" International Journal of Advanced Research in Computer and Software Engineering, Vol 3,Issue 5,May 2013.

[8] M A Anusuya, "Speech recognition by Machine", International Journal of Computer Science and Information security, Vol. 6, number 3,2009.

[9] Sikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad, Arpit Bansal, "Feature Extraction Using MFCC" Signal & Image Processing:An International Journal, Vol 4, No. 4, August 2013.

[10] Mark Gales, Steve Young, "The Application of Hidden Markov Models in Speech Recognition" Foundations and Trends in Signal Processing, Vol 1, No. 3 (2007) 195-304.