Evaluation Framework for Disabled Students based on Speech Recognition Technology

Sanjay Kumar Pal
Department of Computer Science and Applications, NSHM College of Management & Technology,
Kolkata, 700053, India
Email: sarbojay@gmail.com

Seemanta Bhowmick
Department of Computer Science and Applications, NSHM College of Management & Technology,
Kolkata, 700053, India
Email: sam.bhowmick04@gmail.com

Abstract—This paper intends to develop an evaluation framework for the students with disabilities based on speech recognition technology. Education is the most significant ingredient in the development and empowerment of individuals. Till the last decade, education was provided to the persons with disabilities in segregated school settings or "special schools". But in the recent years, there has been a great shift in societal attitude towards disabled students globally. The calls for "integration" of all students, disabled students and non–disabled students into the mainstream classroom environments have gathered momentum worldwide. In the pre–existing frameworks, the disabled students faced great difficulty while interacting with the system. The prime objective of our proposed framework is to provide a user–friendly and interactive environment that gives equal opportunities to all the students being evaluated. The utilization of speech recognition technology would lead to the elimination of all misinterpretations arising due to the human scribe or mediator and would enhance the ability of the disabled students to keep pace with the other students.

Index Terms—Speech Recognition, Hidden Markov Model (HMM), Phoneme, Disabled Students, Evaluation, Framework.

I. INTRODUCTION

"Evaluation is the collection of, analysis, and interpretation of information about any aspect of a programme of education or training as part of a recognised process of judging its effectiveness, its efficiency, and any other outcomes it may have" [17]. Evaluation represents the pinnacle of the learning process, the highest level of learning. The result of evaluation reveals the true depth of understanding of the subject matter by the students.

The judgment of the students' intellectual capability and knowledge can be made primarily by means of evaluation. Thus, the easy availability and accessibility of the evaluation methodology to all spectrums of students becomes vital. The application of the emerging assistive technologies in the field of education has tremendous potential in alleviating the students' problems associated with disabilities [1] [2].

Disability limits access to education and employment due to the lack of necessary support and the means for equal participation. Modern technology could be utilized to assist students with disabilities and provide them with the resources to access the learning opportunities which were previously closed for them and offer them means to perform tasks that they were formerly unable to accomplish, or had great difficulty in accomplishing.

The persons with visual impairment and/or motor impairment are part of the most neglected and excluded sections of the society, and they confront of number of challenges in their day–to–day lives. The mandatory use of keyboard, mouse, or other pointing devices, require eye–hand coordination, and is one of the biggest hurdles for the disabled persons in handling of computers and its associated software.

To help address these problems, the development of an evaluation framework based on the technology of speech recognition is the "need of the hour." The application of speech recognition technology permits hands–free controlling of the framework [9]. The proposed framework would help the visually impaired and motor impaired students on being evaluated without the assistance of any scribe or mediator. The framework would provide an interactive mechanism to "level the playing field" for all the students being evaluated.

II. TERMINOLOGY

A. Speech Recognition

The process by which the computer devices recognizes the spoken inputs by digitizing the sound wave signals

and matching the wave patterns with pre–existing stored wave patterns is referred to as Speech Recognition.

B. Dictionary

The large collection of data used by the speech engine during the process of speech recognition that defines the phonemes in a specified dialect or language is referred to as Dictionary.

C. Grammar

The set of words and phrases that tells the speech recognition software what word patterns to expect in order to perform speech recognition is referred to as Grammar.

D. Natural Language Processing (NLP)

The type of speech recognition approach which tries to determine and understand the intent or meaning of what was being spoken is referred to as Natural Language Processing (NLP).

E. Phoneme

Speech is composed of certain distinct sounds. The smallest unit of speech that conveys a meaning is referred to as Phoneme. It helps us to differentiate one word from another word in a specified language.

To perform speech recognition, the speech engine uses its dictionary to breakup vocabulary words and utterances into phonemes and compares them to one another.

F. Speaker Dependent

The speech recognition software that can recognize the speech input of only the users, it has been trained to understand is referred to as Speaker Dependent.

G. Speaker Independent

The speech recognition software that can recognize the speech input of any user, and does not require any user–specific training is referred to as Speaker Independent.

H. Speech Application

The application in which users interact with the program via means of speech inputs is referred to as Speech Application.

I. Speech Engine

The software that matches the speech inputs to the words and phrases defined in the vocabulary, abiding by the rules specified in the speech recognition grammar is referred to as Speech Engine.

J. Speech Platform

The software that runs the speech application, follows the logic of the application, collects speech inputs, passes the speech inputs to the speech engine, and passes the recognition results back to the application is referred to as Speech Platform.

K. Speech Recognition Grammar Specification (SRGS)

The W3C standards for formulating the grammar of the speech recognition software program are referred to as Speech Recognition Grammar Specification (SRGS).

L. Training

The process of improving the ability of the speech recognizer to adapt to the speakers' voice features and better understand and interpret the speech inputs of the speaker is referred to as Training [13].

M. Utterance

The speaking or vocalization of a word or a set of words, an entire phrase, a sentence, or even multiple sentences is referred to as Utterance.

N. Vocabulary

The complete set of words from which the speech engine tries to make sense of the utterances is referred to as Vocabulary. It comprises of all the words in all active grammars.

III. TIMELINE OF SPEECH RECOGNITION TECHNOLOGY

The development of the speech recognition technology over the past decades has made it possible for the people to control devices and perform other tasks using speech commands. A brief timeline of the evolution of speech recognition technology over the years has been discussed.

- 1922 – The first machine that used the speech recognition technology was a commercial toy produced by Elmwood Button Co. named "Radio Rex." It was a brown bulldog that came out of his doghouse when he heard his name.
- 1928 – The first machine that could generate human speech electronically, when a person entered the words into a special keyboard was called "Vocoder" (acronym for voice coder). It was invented by Homer Dudley at Bell Labs in New Jersey.
- 1952 – The speech recognition system "Audrey", which could recognize only digits spoken by a single user, was designed by the Bell Laboratories.
- 1962 – The "Shoebox" machine was demonstrated by IBM. It could perform simple mathematical calculations via the use of speech commands.
- 1971 – The U.S. Department of Defense's DARPA Speech Understanding Research (SUR) program was responsible for funding Carnegie Mellon's "Harpy" speech–understanding system. Harpy could understand 1011 words, approximately the vocabulary of an average three–year–old.
- 1975 – The description of the speaker–dependent, continuous speech recognition system called the DRAGON system of CMU was laid out by Dr. James Baker. It used uniform stochastic modeling for all sources of knowledge.
- 1985 – The speech–activated typewriter called the "Tangora" was developed by Fred Jelinek. It

contained a vocabulary of 20,000 words and used an IBM PC AT to recognize spoken words and type them on paper.

- 1987 – The continuous speech recognition system called the "BBN BYBLOS" used context–dependent Hidden Markov Model (HMM) to identify the phonemes and initially contained a vocabulary of 997 words [6].
- 1989 – A high performance speech recognition algorithm based on whole word reference patterns for connected digit recognition using Hidden Markov Model (HMM) was developed by Bell Laboratories [3]. It catered to speaker dependent, multi–speaker and speaker independent modes of speech recognition.
- 1990 – A group of continuous speech recognition systems called the "CMU SPHINX" was developed at the Carnegie Mellon University. It included a series of speech recognizers (Sphinx 2–4) and an acoustic model trainer (Sphinx Train).
- 1997 – The company, Dragon came out with diction software, "Dragon NaturallySpeaking" that could recognize human speech and dictate it into a word processing program.
- 1999 – A range of language–specific, speaker dependent, continuous speech recognition system called the "IBM ViaVoice" was offered by IBM.
- 2011 – The intelligent personal assistant, "Siri" was integrated into iPhone 4S and released and has since gone to become an integral part of Apple's products. It uses speech queries to answer questions, make recommendations, and perform actions by delegating requests to a set of Internet service. It even adapts to users' individual language usages, searches, and preferences.
- 2012 – The intelligent personal assistant, "Google Now" was developed by Google. It uses a natural language user interface to proactively deliver information to users that it predicts (based on their search habits) they may want. It is available in the Google app for Android and iOS.
- 2014 – The intelligent personal assistant, "Alexa" was developed by Amazon. It is capable of voice interaction, music playback, making to–do lists, setting alarms, streaming podcasts, playing audio books, and providing weather, traffic, and other real time information, such as news. It can also control several smart devices using itself as a home automation system.
- 2015 – The intelligent personal assistant, "Cortana" was created by Microsoft. It can set reminders, recognize natural voice without the requirement for keyboard input, and answer questions using information from the Bing search engine.

IV. CLASSIFICATION OF SPEECH RECOGNITION

The systems based on the technology of Speech Recognition can be classified into several different classes by describing what the types of utterances they have the ability to recognize, which has been discussed below.

- Isolated Word Recognition – In Isolated Word Recognition, the recognizers require each utterance to have quiet on both sides of the sample window, that is, it require a single utterance at a time. These systems have Listen and Non–Listen states, where they require the speaker to wait or take pause between the utterances.
- Connected Word Recognition – In Connected Word Recognition, the recognizers permit separate utterances to run together with a minimal pause between them. They systems are similar to the Isolated Word Recognition systems.
- Continuous Speech Recognition – In Continuous Speech Recognition, the recognizers permit the users to speak more naturally. It does not require any substantial pause between the words, and it still understands where a word begins and where it ends.
- Spontaneous Speech Recognition – In Spontaneous Speech Recognition, the recognizers have the ability to handle a variety of natural speech features such as "ums" and "ahs" mumblings, and even slight stutters.

V. STAGES OF SPEECH RECOGNITION BASED EVALUATION

Evaluation refers to the process of systematic and continuous determination of a student's merit, worth and significance, using criteria governed by a set of standards. The process of speech recognition based evaluation can be divided into the following stages [12].

Stage 1: Teaching

The "Teaching" stage deals with educating the students, imparting knowledge and sharing of experiences related to the respective disciplines and subjects.

Stage 2: Questions Database

The "Questions Database" stage deals with the preparation of the questions based on the key concepts and subject matter that was taught in the preceding stage, and storing them in appropriate databases for future utilization.

Stage 3: Questions Auto–Generation

The "Question Auto–Generation" stage deals with the automatic generation of the questions that were prepared in the preceding stage in a random order for a fair and uniform examination the students.

Stage 4: Speech Response Gathering

The "Speech Response Gathering" stage deals with the gathering of speech responses of the students, and forwarding them to the succeeding stage for evaluation.

Stage 5: Evaluation

The "Evaluation" stage deals with the evaluation of the speech responses of the students captured in accordance to the questions generated in the preceding stage.

Stage 6: Data Archiving

The "Data Archiving" stage deals with archiving of the data that becomes available after thorough evaluation of the responses gathered in the earlier stages.

Stage 7: Reports Generation

The "Report Generation" stage deals with the generation of appropriate reports based on the facts and figures available after the process of evaluation in the earlier stages.
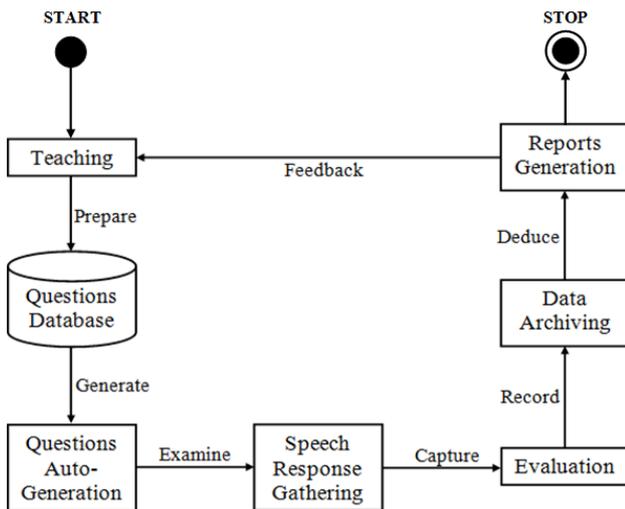

Fig.1. The Seven Stages of Speech Recognition based Evaluation

VI. PROCESS OF SPEECH RECOGNITION

Speech recognition is a process by which the components of the spoken language can be recognized and analyzed, and the semantic message it contains can be translated into a meaningful form. The process of speech recognition can be divided into the following steps [12].

Step 1: Recording of the Speech Input

The user records the speech inputs required by the speech recognizer for the functioning of the framework by speaking into the microphone as shown in Fig. 2.

Step 2: Analogue to Digital Conversion

The sound waves captured by the microphone when the user speaks into the microphone are converted from analogue signals to digitized signals or waveforms as shown in Fig. 3.

Step 3: Removal of the Background Noise

The digitized wave signals are then passed through a signal processing unit called the speech analyzer. The speech analyzer uses Feature Extraction process to remove the background noise from the signals and retain the words that have been spoken into the microphone as shown in Fig. 4 [14].

Step 4: Decomposition into Phonemes

The speech inputs are broken down into phonemes, the smallest unit of sound, by the Speech Analyzer as shown in Fig. 4.

Step 5: Identification of the Phonemes

The speech analyzer passes all the possible phonemes from the speech samples to the Acoustic Model for identification. The Hidden Markov Model (HMM) is used for the task of phoneme identification as shown in Fig. 5.

Step 6: Language Analysis

Once the correct phoneme has been identified, it is then forwarded to the Language Model to determine the correct word that has been spoken and its appropriate meaning as shown in Fig. 6.

Step 7: Generation of the Text Output

After the completion of the analysis conducted by the Language Model, the text output that is the best suited match based on the intricate computations performed is generated as shown in Fig. 6.

Fig.2. Recording of the Speech Input

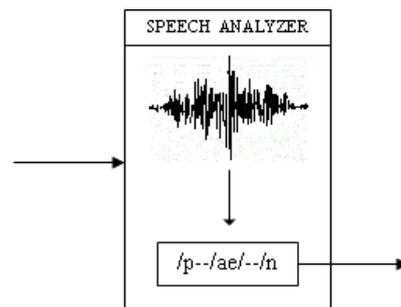Fig.3. Analogue to Digital Conversion

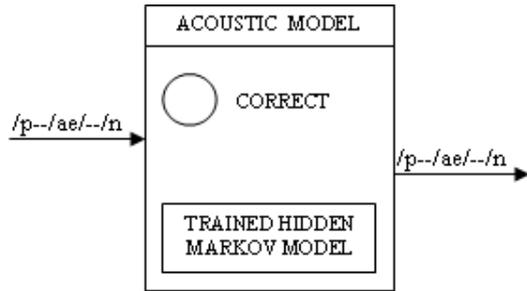Fig.4. Removal of the Background Noise and Decomposition into Phonemes
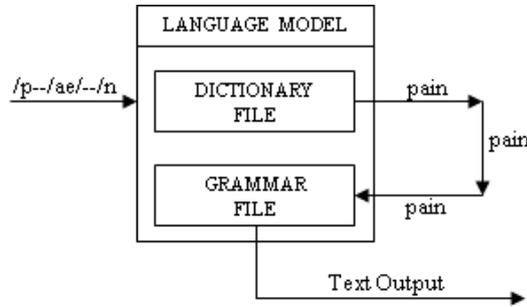
Fig.5. Identification of the Phonemes

Fig.6. Language Analysis and Generation of the Text Output

VII. Using Hidden Markov Model for Speech Recognition

The majority of the modern day speech recognition frameworks are based on the Hidden Markov Model (HMM). The Hidden Markov Model is one of the most important machine learning models in speech and language processing. The HMM is a statistical model in which the framework being modelled is assumed to be a Markov process with hidden states. In a HMM, the states are not directly visible, but the output which is dependent on the states, is visible [5] [7].

The Hidden Markov Model is widely used in speech recognition systems since it permits each speech signal to be viewed as a piecewise immobile and stationary signal or a short–time stationary signal [11]. The speech can be approximated as a stationary wave in a short time–scale period (e.g., 15–20 milliseconds). HMM is also popularly used because it is simple and computationally practical to use and can be trained periodically [15].

A Hidden Markov Model (HMM) can be represented by a 5–tuple (Q, V, π, A, B), where

a) Q is the set of N number of states; these states are unobserved or hidden in the model.

$$Q = \{q_1, q_2, \ldots, q_N\} \qquad (1)$$

b) V is the vocabulary, the set of M number of distinct observation symbols corresponding to the physical output of the framework being modelled.

$$V = \{v_1, v_2, \ldots, v_M\} \qquad (2)$$

c) π is the initial state probability distribution matrix.

$$\pi = \{\pi_i\} \qquad (3)$$

π_i is the probability that the Markov chain will be starting from state i.

$$\pi_i = P[q_1 = i] \qquad (4)$$

$$1 \le i \le N \qquad (5)$$

d) A is the state transition probability distribution matrix.

$$A = \{a_{ij}\} \qquad (6)$$

Each a_{ij} represents the probability of moving from state i to state j.

$$a_{ij} = P[q_t = j \mid q_{t-1} = i] \qquad (7)$$

$$\Sigma a_{ij} = 1 \qquad (8)$$

$$1 \le i, j \le N \qquad (9)$$

e) B is the observation symbol probability distribution matrix.

$$B = \{b_j(k)\} \qquad (10)$$

It expresses the probability of an observation o_t at any instant of time t, being generated from a state i.

$$b_j(k) = P[o_t = v_k \mid q_t = j] \qquad (11)$$

$$1 \le k \le M \qquad (12)$$

VIII. Factors Relevant in Speech Recognition

The working of the speech recognition based systems is affected by various physical and external factors that play a key role in the recognition and subsequent translation of the spoken language. Noise is one of the major environmental factors pertinent in speech recognition. If a speech signal gets infected by noise, the speech quality may be degraded, thereby influencing the process of speech recognition. The most commonly used technique to minimize the effects of background noise is to use a head–mounted microphone. The quality of the captured speech waves depends upon the quality of the digital transducers such as the telephone and microphone. The speech rate varies depending on the different transmission medium, namely, linear medium, bounded medium, uniform medium and isotropic medium. The size of vocabulary and training data available of a speech

recognition system affects the complexity, processing requirements and the accuracy of the system.

The main factors and their respective classification on which speech recognition depends have been presented in the Table 1 [4].

Table 1. Factors Relevant in Speech Recognition

Sl. No.	Factors	Classification
1	Environment	Noise; Type of Noise; Signal–to–Noise Ratio; Working Conditions
2	Transducer	Digital Transducers e.g. Telephone; Microphone
3	Transmission Medium	Linear Medium; Bounded Medium; Uniform Medium; Isotropic Medium
4	System Mode	Speaker Dependent System; Speaker Independent System
5	Speech Pattern	Isolated Word; Connected Word; Continuous Speech; Spontaneous Speech
6	Vocabulary	Discrete Vocabulary; Collective Vocabulary; Training Data Available

IX. Performance Evaluation of Speech Recognition Systems

The performance of speech recognition systems can be evaluated in terms of the speed of recognition and the accuracy of recognition [8] [10].

(a) Speed of Speech Recognition :

The Real Time Factor (RTF) is a measure or metric of the performance of speech recognition systems in terms of speed of the recognition. The RTF is the ratio of the response time to the utterance duration.

$$RTF = \frac{P}{I} \qquad (13)$$

where

- P is the time taken to process the input, and
- I is the duration of the input

(b) Accuracy of Speech Recognition :

The Word Error Rate (WER) is a measure or metric of the performance of speech recognition systems in terms of accuracy of the recognition. The WER is the edit distance between the expected transcription and the system's transcription.

$$WER = \frac{S+D+I}{N} \qquad (14)$$

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions, and
- N is the number of words in the reference.

X. Experimental Results

To test the capability of our proposed framework and evaluate its performance, we conducted two sets of experiments, one under a controlled environment and the other under a natural or uncontrolled environment.

In the first phase of the experiment, we fed the speech recognition system with words in various accents or articulations with the help of a sound emitting device, namely, mobile handset.

This phase of experiment was conducted under controlled conditions, ensuring that there was minimal or no external noise during the experimentation period [16]. The process was repeated ten times for each word in different articulations. And the experiment was carried out for multiple words.

It was observed that when the same word was spoken in different languages and accents, there was a variation in the recognition time of the words which has been shown in Fig. 7. Thus, indicating that recognition of these words is faster when the words are spoken in certain accents or dialects in comparison to others as indicated in Table 2.

Table 2. Variation in Recognition Time with Difference in Articulation

Sl. No.	Articulation (Country)	Recognition Time (in seconds)
1	Bengali (Bangladesh)	0.0084280967712402
2	English (Australia)	0.0073008537292480
3	English (Germany)	0.0076529979705811
4	English (India)	0.0075919628143317
5	English (United Kingdom)	0.0062069892883301
6	English (United States)	0.0069098472595215
7	English (United States, Computer)	0.0071759223937988
8	French (France)	0.0090761184692383
9	Hindi (India)	0.0078909397125244
10	Italian (Italy)	0.0083477828979492
11	Japanese (Japan)	0.0091290473937988
12	Russian (Russia)	0.0086410045623779
13	Spanish (Spain)	0.0087599754333496
14	Thai (Thailand)	0.0092448192596436

In the second phase of the experiment, eighty different speakers (forty male and forty female) were requested to pronounce the eight words from the given list of words in Table 3.

This phase of experiment was conducted under natural or uncontrolled conditions, with no precaution being taken for ensuring that there was minimal or no external noise during the experiment. The process was repeated eighty times for each word by different speakers, both male and female. And the experiment was carried out for multiple words.

It was observed that some words had a higher recognition rate when compared to others. The variation in the recognition rate of these words has been shown in Fig. 8. The experiment was repeated for each word and the corresponding recognition rate of the words was noted down. The average recognition rate of these words in percentage is given in Table 3.

Table 3. Variation in Recognition Rate for the Different Speech Commands

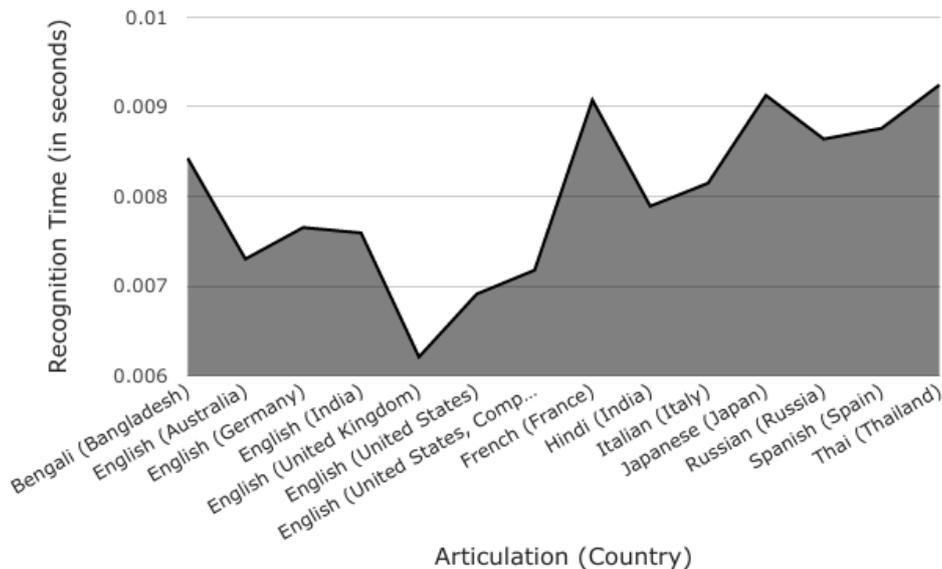| Sl. No.	Speech Commands	Recognition Rate (in percentage)
1	Login	97.8%
2	New User Registration	93.6%
3	Start Exam	98.8%
4	Move Forward	96.2%
5	Move Backward	94.5%
6	Flag Question	95.6%
7	Submit Exam	98.4%
8	Logout	97.6%

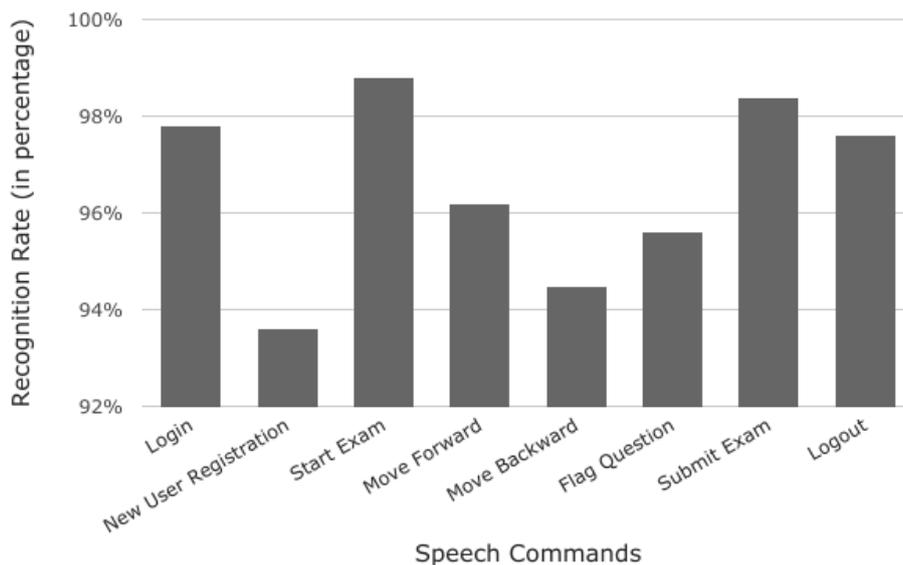Fig.7.Chart depicting the Variation in Recognition Time

Fig.8. Chart depicting the Variation in Recognition Rate

XI. CONCLUSION

The rapid emergence of speech recognition technology has captured the imagination of millions of people worldwide. The broad potential of the speech recognition technology to engage, interact and serve in a more effective way has fuelled the fire among researchers to utilize this technology and venture into new sectors. The modern advancements in speech recognition technology shows a lot of promise and capability to further bridge the gap between the human and machine interactions.

REFERENCES

[1] B.H. Juang and Lawrence R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development".

[2] Kirsi Jääskeläinen and Nina Nevala, "Use of Assistive Technology in Workplaces of Employees with Physical and Cognitive Disabilities", K. Miesenberger et al. (Eds.): ICCHP 2012, Part I, LNCS 7382, pp. 223–226, 2012.

[3] Lawrence R. Rabiner, Jay G. Wilpon and Frank K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models", IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, Vol. 37, Issue 8, August 1989.

[4] Shikha Gupta, Mr. Amit Pathak, Mr. Achal Saraf, "A STUDY ON SPEECH RECOGNITION SYSTEM : A LITERATURE REVIEW", International Journal of Science, Engineering and Technology Research, Vol. 3, Issue 8, August 2014.

[5] B.H. Juang and Lawrence R. Rabiner, "Hidden Markov Models for Speech Recognition", Technometrics, Vol. 33, No. 3, pp. 251–272, August 1991.

[6] Richard Schwartz, Chris Barry, Yen–Lu Chow et al., "The BBN BYBLOS Continuous Speech Recognition System."

[7] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", PROCEEDINGS OF THE IEEE, Vol. 77, No. 2, pp. 21–27, February 1989.

[8] Suma Swamy and K.V. Ramakrishnan, "An Efficient Speech Recognition System", Computer Science & Engineering Journal, Vol. 3, No. 4, August 2013.

[9] Fouzia Khursheed Ahmad, "Use of Assistive Technology in Inclusive Education: Making Room for Diverse Learning Needs", Transcience, Vol. 6, Issue 2, pp. 62–77, 2015.

[10] Preeti Saini and Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology, Vol. 2, Issue 11, pp. 132–136, November 2014.

[11] Sumalatha. V and Dr. Santhi. R., "A Study on Hidden Markov Model (HMM)", International Journal of Advance Research in Computer Science and Management Studies, Vol. 4, Issue 2, pp. 465–469, 2013.

[12] Suma Shankaranand, Manasa S, Mani Sharma, et al., "An Enhanced Speech Recognition System", International Journal of Recent Development in Engineering and Technology, Vol. 2, Issue 3, pp. 78–81, March 2014.

[13] Satyanand Singh, Abhay Kumar, David Raju Kolluri, "Efficient Modelling Technique based Speaker Recognition under Limited Speech Data", International Journal of Image, Graphics and Signal Processing (IJIGSP), Vol.8, No.11, pp.41–48, 2016. DOI: 10.5815/ijigsp.2016.11.06

[14] Anjali Pahwa, Gaurav Aggarwal, "Speech Feature Extraction for Gender Recognition", International Journal of Image, Graphics and Signal Processing (IJIGSP), Vol.8, No.9, pp.17–25, 2016. DOI: 10.5815/ijigsp.2016.09.03

[15] Hajer Rahali, Zied Hajaiej, Noureddine Ellouze, "Robust Features for Speech Recognition using Temporal Filtering Technique in the Presence of Impulsive Noise", IJIGSP, Vol.6, No.11, pp.17–24, 2014. DOI: 10.5815/ijigsp.2014.11.03

[16] Ravi Kumar. K, P.V. Subbaiah, "A Survey on Speech Enhancement Methodologies", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.12, pp.37–45, 2016. DOI: 10.5815/ijisa.2016.12.05

[17] Mary Thorpe, "Handbook of Education Technology", Ellington, Percival and Race, 1988.

Authors' Profiles

Sanjay Kumar Pal is working as an Associate Professor in NSHM College of Management & Technology, Kolkata. He has completed MCA & M.Tech (Information Technology), and has already presented his Doctoral Public Seminar on "Fascicles of Graph Algorithms". He has over 24 years of experience shared between 11 years in Industry and 13 years in teaching. He has published a book on Graph Theory, "Allurement of Some Graph Algorithms" and more than 50 Research Papers in different National and International Journals.

Seemanta Bhowmick is a final year Masters of Computer Science student in NSHM College of Management & Technology, Kolkata. He has completed his Bachelors in Computer Applications and the Course Topper award was conferred upon him in 2016. Recently, he was adjudged the Best Volunteer in the Vittiya Saksharta Abhiyan, a Digital Financial Literacy Campaign in 2017 and was duly felicitated by the Honorable Minister of Human Resource Development, Government of India.