International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

# Voice Recognition System Through Machine Learning

**Sindhu B, B Sujatha**

***Abstract*: Human voice recognition by computers has been ever developing area since 1952. It is challenging task for a computer to understand and act according to human voice rather than to commands or programs. The reason is that no two human's voice or style or pitch will be similar and every word is not pronounced by everyone in a similar fashion. Background noises and disturbances may confuse the system. The voice or accent of the same person may change according to the user's mood, situation, time etc. despite of all these challenges, voice recognition and speech to text conversion has reached a successful stage. Voice processing technology deserves still more research. As a tip of iceberg of this research we contribute our work on this are and we propose a new method i.e., VRSML (Voice Recognition System through Machine Learning) mainly focuses on Speech to text conversion, then analyzing the text extracted from speech in the form of tokens through Machine Learning. After analyzing the derived text, reports are created in textual as well graphical format to represent the vocabulary levels used in that speech. As Supervised learning algorithm from Machine Learning is employed to classify the tokens derived from text, the reports will be more accurate and will be generated faster.***

***Keywords : Machine Learning, Speech Recognition, Speech to text conversion, Text analysis, Vocabulary***

## I. INTRODUCTION

Voice-recognition software programmes work by analysing sounds and changing them to text. They additionally use information of how English is typically spoken to come to a decision what the speaker most likely spoke. Once properly established, the systems ought to recognise around ninety fifth of what's aforementioned if you speak clearly. Many programmes are on the market that offers voice recognition. These systems have principally been designed for Windows operating systems.

Speech is the primary mode of communication between individuals.

Speech recognition, generation of speech waveforms, has been facing least development for many decades [1]. Such speech contains lots of information. Most prominent and common mode of information transfer is through speech. In this system we are going to extract text from the speech and then vocabulary from that text to analyse the user's vocabulary skill set.

**Manuscript published on 30 August 2019.**
***Correspondence Author(s)**
**Sindhu. B**, Department of Computer Science and Engineering,, Godavari Institute of Engineering and Technology, Rajamahendravaram, Andhra Pradesh, India
**Dr B Sujatha**, Department of Computer Science and Engineering,, Godavari Institute of Engineering and Technology, Rajamahendravaram, Andhra Pradesh, India



## II. LITERATURE SURVEY

Automatic speech Recognition is a method by that a PC takes a speech signal and converts it into words [2]. It is the method by which a PC acknowledges what a person said. Keyboard, though a standard medium, is not terribly convenient, as it needs a bound quantity of talent for effective usage .A mouse on the different hand needs a sensible hand eye co-ordination. Physically challenged individuals notice PC tough to use. Partly blind individuals notice reading from a monitor tough. All these constraints have to be eliminated. Speech interface facilitate us to tackle these issues. The objective is to entice human voice during a electronic computer and decipher it into corresponding text. Speech recognition will be outlined as the method of changing an acoustic signal, captured by a electro-acoustic transducer (microphone) or a telephone, to a group of words. When two individuals speak to one another, they each acknowledge the words and will be able to understand the meaning behind them. Computers, on the opposite hand, are solely capable of the initial thing: they can only acknowledge individual words and phrases; however they don't extremely perceive speech within the same means as humans do. PC acknowledges the command and software system tells the PC what to try to once that command is recognized [3].

Big data refers to information volumes within an extent of Exabyte ($10^{18}$ B) and on the far side. Such volumes exceed the capability of current on-line storage and process systems. With characteristics like volume, velocity and variety massive information throws challenges to the normal IT sector. Computer aided innovation, real time information analytics, customer-centric business intelligence; trade wide higher cognitive process and transparency are potential blessings, to say few, of huge information. There are several problems with massive information that warrant quality assessment strategies. The problems are touching on storage and transport, management, and process. This paper throws lightweight into this state of quality problems associated with massive information. It provides valuable insights that may be accustomed leverage massive information science activities. [13] Big data and its analysis are at the middle of contemporary science and business. These information are generated from on-line transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science information, sensors and mobile phones and their applications [4][5]. They are hold on in databases grow massively and become troublesome to capture, form, store, manage, share,

analyze and visualize via typical information software system tools. 5 exa-bytes (1018 bytes) of information were created by human till 2003. Nowadays this quantity of knowledge is formed in 2 days. In 2012, digital world of information was enlarged to 2.72 zetta-bytes (1021 bytes). It's foretold to double each 2 years, reaching regarding eight zetta-bytes of information by 2015 [6]. IBM indicates that each day 2.5 exa-bytes of information created additionally ninetieth of the information created in last 2 years [7]. A PC holds concerning five hundred gigabytes (109 bytes), thus it would need concerning twenty billion computers to store all of the world's information. Within the past, human genome decryption method takes about ten years, no longer quite every week [8]. Multimedia information have huge weight on web backbone traffic and is expected to increase seventieth by 2013[9]. Solely Google has got additional than one million servers around the world. There have been six billion mobile subscriptions in the world and each day ten billion text messages are sent. By the year 2020, fifty billion devices are connected to networks and therefore the web [10]. In 2012, The Human Face of huge information accomplished as a worldwide project, that is centring in real time collect, visualize and analyze massive amounts of information. According to this media project several statistics are derived. Facebook has 955 million monthly active accounts with exploitation of seventy languages, a hundred and forty billion photos uploaded, one hundred twenty five billion friend connections, each day thirty billion items of content and 2.7 billion likes and comments are announced. Each minute, forty eight hours of video are uploaded and each day, four billion views performed on YouTube. Google support several services as each monitors 7.2 billion pages per day and processes 20 peta bytes (1015 bytes) of information daily additionally interprets into sixty six languages. One billion Tweets each seventy two hours from additional than a hundred and forty million active users on Twitter. 571 new websites are created each minute of the day [11]. Among consequent decade, range of data can increase by fifty times but range of data technology specialists who continue with all that information can increase by 1.5 times. [12].

Machine Learning develops a formula supporting the data analysis as now-a-days great amount of information is accessible all over. Therefore, it's vital to associate this knowledge so as to extract some helpful data. This may be achieved through data processing and Machine Learning. Machine learning is an integral a part of computing, that is employed to create algorithms supported by the information trends and historical relationships between data. Machine learning is employed in varied fields like bioinformatics, intrusion detection, data retrieval, game enjoying, marketing, malware detection, image de-convolution and then on.[14] Machine Learning (ML) will be thought of as a subfield of Artificial Intelligence since those algorithms will be seen as building blocks to build computers learn to behave intelligently by somehow generalizing rather than simply storing and retrieving knowledge things like a information system and alternative applications would do. Machine learning has got its inspiration from a sort of educational and academic disciplines, together with applied science, statistics, biology, and science. The core functionality of Machine learning makes an attempt is to tell computers however to mechanically realize a smart predictor primarily based on past experiences and this job is done by smart classifier. Classification is the method of exploitation a model to predict unknown values (output variables), using a range of well-known values (input variables). [15]

*A. Machine Learning Algorithms*

Machine learning algorithms are the programmes which will continuously learn from the given knowledge and improve from gained experience, while there is no human intervention needed. Learning tasks will include learning a mathematical function that maps input to the output, then the system learns the hidden structure in data that is not labelled; or 'instance-based learning', where a class label is provided for every new instance by rapidly comparing the fresh instance (row) to instances from the existing training data, which were stored in memory. [16] Machine learning algorithms are divided into three categories according to the feature how they learn. They are as follows

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning [17]

Supervised Learning: Supervised learning as itself the name indicates the presence of a supervisor as an educator. Essentially supervised learning may be a learning during which we have a tendency to teach or train the machine using knowledge which is well labelled which means some knowledge is already labelled with the proper answer. After that, the machine is supplied with a brand new set of examples (data) in order that supervised learning formula analyses the coaching knowledge (set of coaching examples) and produces an accurate outcome from labelled data. Supervised Learning algorithms are Regression, Decision Tree, Random Forest, KNN, Logistic Regression algorithms etc . [18]

Classification is employed to predict the end result of a given sample once the output variable is within the style of classes. A classification model would possibly check out the input data and take a look at to predict labels like "medium" or "heavy." [19]

Regression is employed to predict the end result of a given sample once the output variable is within the style of real values. For instance, a regression model would possibly process input data to predict the number of downfall, the peak of someone, etc. [20]

Ensembling is another sort of supervised learning. It suggests that combining the predictions of multiple machine learning models that square measure severally weak to provide a lot of correct prediction on a replacement sample.

Unsupervised Learning: In this approach, we do not have any outcome variable to predict the situation. This is used for mainly clustering the population into different groups, where it is very widely used for customer segmentation into different groups for specific intervention or purposes. Examples of Unsupervised Learning algorithms are



Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Apriori algorithm, K-means clustering algorithms

Association is employed to get the chance of the co-occurrence of things during a collection. It's extensively utilized in market-basket analysis. For instance, AN association model could be accustomed discover that if a client purchases bread, he is 80th probably to additionally purchase eggs.

Clustering is employed to cluster samples such objects at intervals constant cluster are additional the same as one another than to the objects from another cluster.

Dimensionality Reduction is employed to cut back the quantity of variables of a knowledge set whereas guaranteeing that necessary information continues to be sent. Dimensionality Reduction will be done using Feature Extraction ways and have choice methods. Feature selection selects a set of the first variables. Feature Extraction performs information transformation from a high-dimensional space to a low-dimensional space. [21]

Reinforcement Learning: In this approach, the machine is trained to create specific choices. It works this way: the machine is exposed to Associate in Nursing atmosphere wherever it trains itself frequently using trial and error. This machine learns from past expertise and tries to capture the simplest attainable information to create correct business choices. Example of Reinforcement Learning: Markov decision method. The following diagram represents methodology of reinforcement learning.

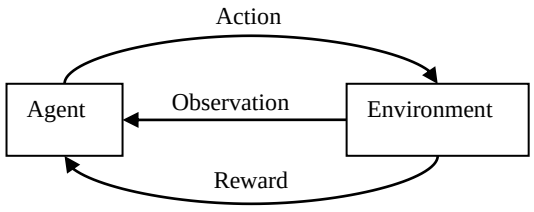


**Fig. 1. Methodology of reinforcement learning**

*B. K- Nearest Neighbour Algorithm*

The K-Nearest Neighbour algorithm is a non-parametric and the most powerful and easiest algorithm amongst all the machine learning algorithms. It is a category of instance-based learning, or even lazy learning, because the calculation function is calculated only locally and all computation is deferred until classification is done. It uses the entire data set (called as feature space) as the training set, rather than splitting the data set into a training set and test set. When an outcome is required for a new data instance, the KNN algorithm goes through the entire data set to find the k-nearest instances to the new instance. The value of k is user-specified . The similarity between two instances is calculated using quantitative measures such as Euclidean distance or Hamming distance. These are often used for every classification and regression issue.

However, it's a lot of widely utilized in classification issues within the business.

K nearest neighbours may be a straightforward rule that stores all offered cases and classifies new cases by a majority vote of its k neighbours. The case being appointed to the category is commonest amongst its K nearest neighbours measured by a distance operate. [22]

*C. Algorithm to implement K-Nearest Neighbour algorithm*

1) Load the whole data
2) Initialise value for k
3) For deriving the output (predicted class), start iterating from first to end training data points

a. Calculate the distance between test data and each and every row of training data. Use Euclidean distance as the distance metric. Other metrics Chebyshev, cosine, etc can also be used as needed.
b. Based on the distance values, sort the calculated distance in ascending format
c. Sort the calculated distances in ascending order based on distance values
d. Extract the top k rows from the sorted array
e. Extract the most frequent class known of these rows
f. Return the output (predicted class)

## III. METHODOLOGY

*A. Proposed Method*

The process of proposed VRSML (Voice Recognition System through Machine Learning) is represented in the form of a flowchart as below. Voice which is given as the input is parallelly recorded as audio file and converted to text. Then the text is tokenized and passed to the analyzer, all the tokens are then categorized based on GSL (General Service List) by applying K-Nearest Neighbour algorithm of Machine Learning, finally a report is generated which describes the vocabulary used in the input speech.

Below is the detailed step by step procedure followed by VRSML

Step 1: An input is given to the system in the form of English speech which is generated by humans using microphones connected to the computer. The input can be either taken from Big data.

Step 2: The human generated audio is recorded by the recorder as and then the user speaks. It can record the audio for unlimited amount of time.

Step 3: The recorded audio is saved in .mp3 format for any further references and cross-checking the converted text.

Step4: Convert the generated audio into text format using Speech to Text Converter tool.

Step 5: Save the extracted text into a .txt file format. These files are further processed to extract the level of vocabulary used.

Step 6: The whole text is given as input to a tokenizer and it tokenizes the whole text and extracts each and every single word from it accepting the character space as delimiter.

Step 7: Using Recurrent Neural Networks we make the system learn from the given inputs and make an appropriate decision.

Step 8: Using K-Nearest Neighbour algorithm of supervised learning approach of Machine Learning, the extracted tokens are classified into several categories such as repeated or not, general word or not. If repeated, then for how many times it was repeated and the number of words that are been used uniquely. In brief, a summary of the vocabulary levels used in the speech is generated.

Step 9: Textual summary as well graphical summary (Point Graph, Bar Graph) are used to show the vocabulary levels used in the speech generated.

Step 10: With the use of the summary generated by the system, one can analyze their own vocabulary skills or can assess others as well

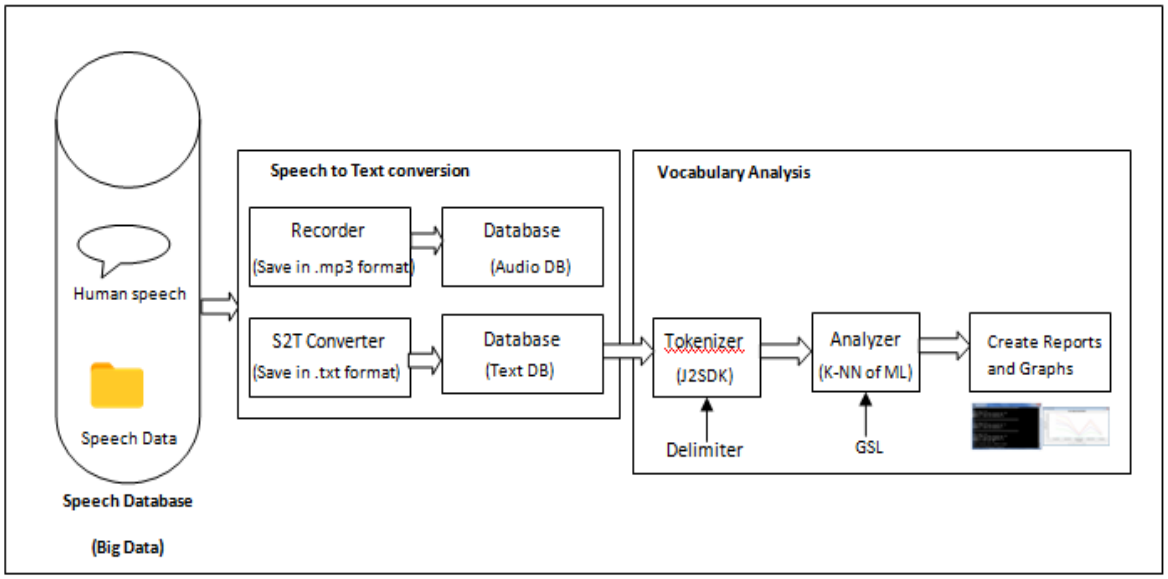


**Fig. 2. Block diagram of VRSML**

*B. Building process:*

In the process of completing the project successfully, we have gone through some crucial phases of it. Those are described below.

- Accepting, recording and saving human generated audio through a connected microphone (Speech Recognition)
- Converting the audio into text and saving it (Speech to Text Conversion)
- Converting the text into tokens (Tokenization)
- Analyze and categorize the tokens using Supervised Learning with help of GSL (Machine Learning)
- Creating reports (UI)

*C. Existing system vs proposed system*

Till date there are no such system which performs both speech to text conversion and vocabulary extraction as well from the given speech. Hence there does not exist any existing system. In proposed system we have incorporated both the functionalities in a single system.

## IV. EXPERIMENTAL RESULTS

*A. Performance measurement:*

The performance of this application is very effective and efficient. We have stringently tested this application under critical and crucial inputs. This application performs very efficiently for every input and did not fail to work in any aspect. It could be said that this specific application is 99.99% accurate as it can handle any valid input and generate accurate report.

**Table 1: Performance measurement of the system**

| Number of speeches | 100 | 600 | 1500 |
|---|---|---|---|
| Speech Recording Accuracy rate | 100% | 100% | 100% |
| Speech Conversion Accuracy rate | 99% | 99% | 99% |
| Text Analysis Performance rate | 100% | 100% | 100% |
| Accuracy Rate Of Graphical Representation | 100% | 100% | 100% |
| Execution time in milli seconds | 12ms | 24ms | 30ms |

*B. BLEU SCORE*

BLEU score refers to Bilingual Evaluation Understudy Score which is a metric used in NLP (Natural Language Processing) to rate the text generated by a system which is cross checked with one or more reference translations given by humans.
This metric is used where output is a text string rather than being a classification. It is an algorithm for evaluating the performance of machine translation. This is one of the most popular metric which tells us how good the translation system is working and comes closer to human translation. This works with uni-gram, bi-gram, tri-gram, 4-gram till n-gram comparison. In uni-gram approach of comparison, look at each word in the sentence and then assign a score of 1 if it is included in any of the given reference sentences. Assign a 0 if it is not included in any of the reference strings. Divide the derived sum of values with total number of words given in the reference string. This gives the unigram precision value. The maximum score is 1 and 100 in terms of percentage, which determinates that the system is working perfect. Following is the formula to calculate BLEU score.

$$BLEU = \min\left(\left(1, \frac{output-length}{reference-length}\right), \left(\prod_{i=1}^{4} precision_i\right)^{1/4}\right)$$

We can also examine n- sub sequent words that occur in a reference string that is n-gram approach. This method does not have any concern with meaning of the text or with the sentence structure but only compares the generated string with the reference strings. We have tested our application with BLEU score and it has scored **98%.** So it can be said that this application performs very well in speech to text conversion in every scenario.

*C. Observations*

The observations made during the execution of proposed system are that only general speech is recognized. Expressions or emotions of the user are not considered. Punctuations given in the speech are not considered. Space is evaluated only as a delimiter but not as a character. Speech of a single user is only considered. Recognizer will only recognize the voice generated at a diameter of 2.5cms of the connected microphone.

*D. Result analysis*

The results are analyzed and following statements are made during the execution of this tool. We have prepared audio data set consists of 100 .mp3 files. All the audios are converted to text and saved as .txt files. The recorder www.recordermp3online.com is used to record the given audio and save it to the computer in .mp3 format. .mp3 format is used as files encoded in MP3 have a quality very similar to that of the CD audio tracks but are much smaller in size. MP3 files are around 11 times smaller than uncompressed music tracks. The recording done by the recorder will be helpful to manually check the converted text. The converter www.speechnotes.co is used accept input from the user in the form of voice. Text analyzer logic is implemented in Java programming language which accepts the converted text file and analyzes it. The output of Text Analyzer consists of the following categories

- Total number of words extracted
- Number of words used uniquely
- Number of words used repeated
- Number of general words used
- Number of non-general words used
- Summary of repeated words and its count

Here we have summarized some of the major differences and difficulties faced in manual and computerized interpretation of human speech. In manual interpretation multiple voices can be recognized where it cannot be done through computerized recognition. Distortion, pauses, delays in speech cannot be understood by computerized interpretation but it can be done by humans. Text analysis, Speech to Text conversion, Text interpretation accuracy depends on the dexterity of the person who handles it but in the case of computerized interpretation all these will be very accurate and depends on the training given to the system.

In the following table summary of differences occurred between normal classification and K-Nearest Neighbour classification of tokens is realized.

**Table 2. Normal classification of tokens Vs K-NN classification**

| Scenario | Normal Classification approach | K- NN classification approach |
|---|---|---|
| Accuracy | 99% | 100% |
| Speed | Moderate | Fast |
| Approach | Non-intelligent | Intelligent |
| Mode | Non-Predictive | Predictive |
| Learning approach | NA | Supervised Learning |
| Categorization approach | Only iterative comparisons | Based on the knowledge gained in previous cases |

*E. Screens*

Following are screens captured during execution of the system

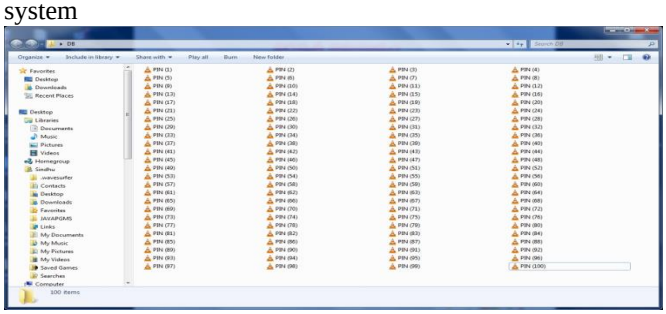

**Fig. 3. Audio data set of 100 voice samples**

**Fig 4. Text data sent which is derived by converting Speech to Text conversion**

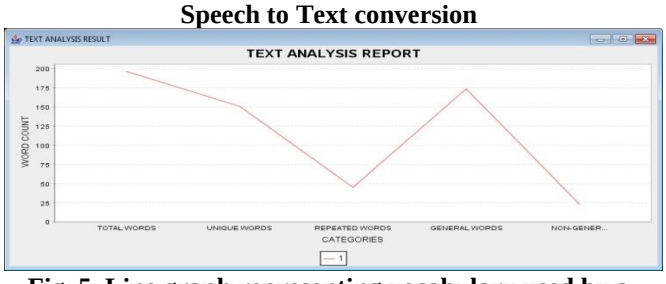


**Fig. 5. Line graph representing vocabulary used by a single user**

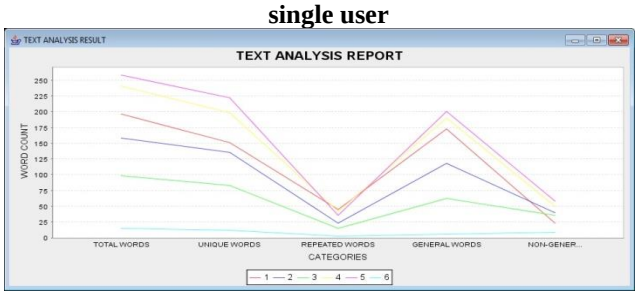


**Fig. 6. Line graph representing vocabulary used by 6 users**

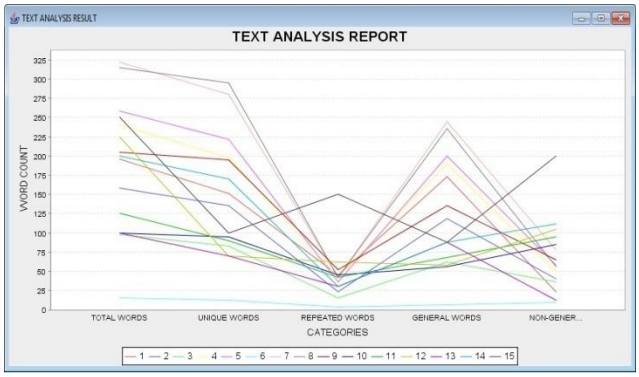


**Fig. 7. Line graph representing vocabulary used by 15 users**

## V. CONCLUSION AND FUTURE WORK

We strongly believe that communication skills are very next to technical skills. If one is blessed with plenty of technical skills but is not able to communicate with others, he will be surely not considered to be skilled enough. Vocabulary skills occupy major part of communication skills. So vocabulary dexterity plays a vital role for every student or a techie's life. Analyzing such an important skill shows technical advancements in computer technologies. This application is developed keeping in mind for people belonging every aspect. But the only drawback is that it works for only English language. In future, it is suggested to upgrade this application to work with maximum possible national and regional languages and make our nation as well the whole world more strengthened in communication and be proficient enough.

*A. Scope of the project:*

This project has very wide scope as it can analyze any user's vocabulary levels within fraction of seconds. It can be used in any sort of educational institutions, it can be used from primary schools till universities to assess their student's or faculty's dexterity in English Vocabulary and act accordingly to improve their vocabulary skill set. This tool can be used in corporate offices where English communication plays a vital role. This tool also helps in interview phases and makes the work of interviewer to analyze the vocabulary levels of the interviewee done faster. This can even be used for self assessment of one's vocabulary levels. In brief there is a wide scope for this project wherever there is a scope for communication in English.

## REFERENCES


1. Rabiner Lawrence ,Juang Biing-Hwang, "Fundamental of speech recognition ",AT & T,1993.
2. Anne Johnstone Department of Artificial Intelligence Edinburgh University Hope Park Square, Meadow Lane Edinburgh EHB 9LL, (GB) Gerry Altmann "Automated speech recognition:a framework for research".
3. IJARCCE www.ijarrcce.com Speech recognition system for english languagems. Vrinda ,Mr. Chander Shekhar Astt.Professor,Deptt of Computer Science& Engineering, M. V N University,Palwal, India
4. C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012
5. R.D. Schneider, Hadoop for Dummies Special Edition, John Wiley&Sons Canada, 978-1-118-25051-8, 2012
6. Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012
7. S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011
8. http://en.wikipedia.org/wiki/Big_data , last access 11.03.2013
9. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", mckinsey Global Institute,2011,
10. B.Gerhardt, K. Griffin and R. Klemann, "Unlocking Value in the Fragmented World of Big Data Analytics", Cisco Internet Business Solutions Group, June 2012,
11. http://www.humanfaceofbigdata.com/ , last access 11.03.2013
12. C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012
13. G. A. Lakshen, S. Vraneš and V. Janev, "Big data and quality: A literature review," 2016 24th Telecommunications Forum (TELFOR), Belgrade, 2016, pp. 1-4. http://ieeexplore.ieee.org/stamp/stamp.jsp?Tp=&arnumber=7818902 & isnumber=7818703
14. S. Angra and S. Ahuja, "Machine learning and its applications: A review," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala,2017,pp.57-60.
15. Supervised machine learning approaches: a survey iqbal muhammad1and zhu yan2school of Information Sciences and Technology, Southwest Jiaotong University, China
16. https://www.edureka.co/blog/machine-learning-algorithms/
17. Qbal muhammad and zhu yan: Supervised machine learning approaches: a survey doi: 10.21917/ijsc.2015.0133
18. https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14
19. https://machinelearningmastery.com/confusion-matrix-machine-learning/
20. https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html
21. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
22. https://www.geeksforgeeks.org/k-nearest-neighbours/