Speech Recognition System: A review

Article *in* International Journal of Future Generation Communication and Networking · July 2020

3 authors, including:

Sachin Sharma
Manav Rachna Educational Institutions
15 PUBLICATIONS **17** CITATIONS

SEE PROFILE

International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 2547–2559

Speech Recognition System: A review

Rohit Pahwa, Harion Tanwar, Dr Sachin Sharma
Manav Rachna International Institute of Research and Studies, Faridabad, Haryana

Abstract:
Speech recognition system play an essential role in every human being life. It is a software that allows the user to interact with their mobile phones through speech. Speech recognition software splitting down the audio of a speech into various sound waves forms, analyzing each sound form, using various algorithms to find the most appropriate word fit in that language and transcribing those sounds into text. This paper will illustrate the popular existing system namely SIRI, CORTANA, GOOGLE ASSISTANT, ALEXA, BIXBY. Apart from that, this paper also analysis the concept of NLP (Natural processing) with speech recognition. In addition to this, our main function is to find out the most accurate technique of speech recognition so that we can achieve the best results. Comparative analysis will indicate the difference and demerit points of various speech recognition.

Keywords: *Speech recognition, Deep learning, SIRI, CORTANA, ALEXA, BIXBY, GOOGLE ASSISTANT, Acoustic Models, NLP*

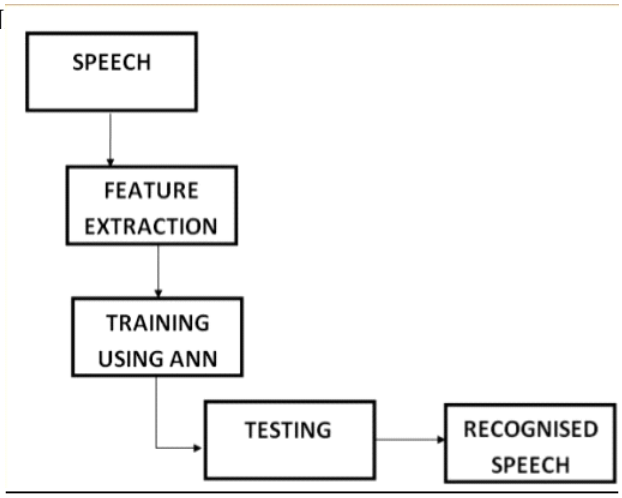
1. Introduction

Speech Recognition is a kind of technology which allows the user to operate the electronic device through spoken word instead of using different tools such as keystrokes, button and keyboard etc. Speech recognition software convert the words and phrases which is spoken by user into machine-readable format so that user can easily operate the device through speech. Speech recognition which is also known as automatic speech recognition (ASR). The main objective of developing speech recognition is any people whether it is technical or non-technical can easily operate the device. As well as an illiterate which have no knowledge about device and its parts they can be operated it very easily. Speech recognition is basically designed for a single user. The block model of Speech recognition system is shown in Figure 1. The field of speech recognition is an emerging research area with important application in Banking, marketing, healthcare, language learning and many more. There are various parameter in speech processing such as pitch, duration, voice quality, intensity, signal-to-noise ratio, voice activity detection and strength of voice. What speech recognition does it

extract all this parameter from the algorithm so that it can execute the user query on the behalf of these parameters. Nowadays, speech

recognition is working with various technology such as m

achine learning, IOT,N



ISSN: 2233-7857 IJFGCN

Figure 1: Speech Recognition System

[8]

The main motive of speech recognition is WHO is speaking and WHAT was spoken. It is used to identify a person by analyzing its tone, voice pitch and many more parameter.[8]

Voice recognition was developed as a faster method of typing up work and was originally designed with people who suffer from various disabilities in mind as well as people who suffer with physical difficulties can find typing tedious, painful or even impossible and this give them the chance to still do in the manner of speaking. Voice recognition

software allows you to dictate to your computer. When you will give the instruction through this software on that moment all the instruction will be appeared on the screen. It is so advanced that it can predict what you want to say, so that it can correct error of mis-speech or grammatical errors for you.

Working of Speech Recognition process

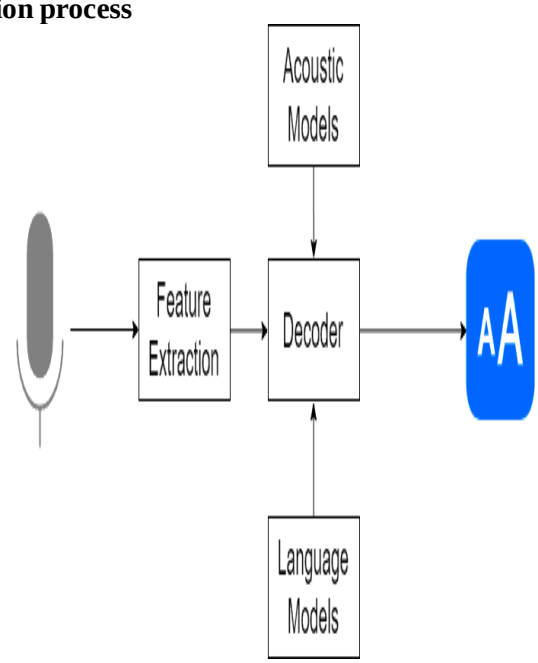


Figure 2: Working of Speech Recognition process[10]

1) Feature Extraction:-

Speech recognition software will analysis the sound through various parameter such as pitch, variation, strength of voice.

2) Acoustic Models:

This model is a computational file that contain different forms of sound that make up a word. It shows the relationship between audio signal and other basic language unit that create the speech. These computational and mathematical representation is known as HMM's model.

3) Decoder:

Neural network break down the speech in various neurons. we have a different types of algorithm they will

ISSN: 2233-7857 IJFGCN

decode these neurons through these algorithm so that we can get the correct result.

4) Language Model:

The language model in speech recognition helps differentiate between words and phrases that sound similar .When we integrate the evidence of language model with pronunciation and acoustic model, with the help of this we can solve the problem of ambiguity. [9]

The remaining paper is prepared as follow: Section 2 describes the various speech recognition software. Section 3 machine learning with speech recognition. Section 4 contain the comparative analysis of different speech recognition. Section 5 summarizes conclusion and the tail of the paper contain conclusion.

Various Speech Recognition system

Speech recognition system is a machine or program that helps to identify words and phrases in spoken audio and convert them into a machine understandable format but various speech recognition software has a limited vocabulary of words and phrase and it may only identify these words which are available in their database or which they are spoken by very clearly.

Existing System which uses speech recognition

- **APPLE (SIRI):**

Siri is a virtual assistant is a part of Apple Inc. It is designed to offer you a multiple way of interaction with your phone by speak up. It will take the query through microphone and help you solve the query with in time. It has some features that make it different from other speech recognition, for instance, can activate low power mode, Enable do-not disturb mode as well it has Non-English option. Most of them use it as entertaining purpose. On the other hand, Siri has one demerit point it only works on IOS devices.[4]

- **GOOGLE ASSISTANT:**

Google Assistant is virtual assistant of google Inc's. Google assistant control your devices and smart phone. Some important features of Google assistant are, it control your device and your smartphone and access information from calendar. As well as it can also handle your music system. It has some demerit points such as it use more battery power due to this sometime it slow down the working of system.[6]

- **MICROSOFT CORTANA**

Like Siri and Google Assistant, Cortana is also a voice assistant developed and created by Microsoft. Basically it is designed for window devices Nowadays, it is available in various devices .It can performs a multiple task for users, like remainder setting, as well as it can also scheduling the calendar events, even most of people who use Cortana to performing some computational data and many more Cortana has an API (application programming interface) and can work with a variety of windows app, as well as third-party apps such as Facebook and Twitter. Apart from that, it has several demerits such as Vulnerability found hit the listening button again and again and many more.

- **AMAZON ALEXA**

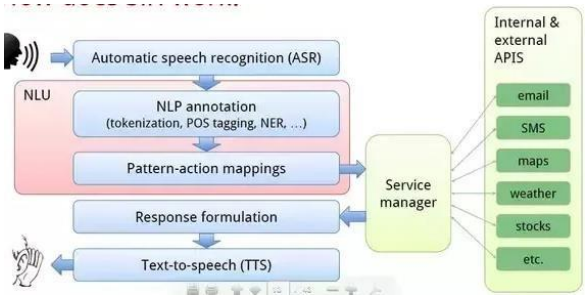
Alexa is a virtual assistant technology designed and created by Amazon. This technology is based upon Machine learning, NLP(natural language processing).Alexa can perform various task such as it can acknowledge the user about Weather. Furthermore, it can handle your smart phone like when user give the instruction it will take and solve it with in time so that user can do another work. But it has several demerit points, for example, it cannot send a message and Email through voice command. Apart from that, if a person wants to access the heath and hobbies related data through Alexa, in that cases Alexa will not give the accurate result.[1]

ISSN: 2233-7857 IJFGCN

- **SAMSUNG BIXBY**

Bixby is a virtual assistant developed by Samsung Electronics. With the help of this system you can send message from one device to another ,as well as you can check the cricket and any other game score. A hardware button on the side of the device to bring up Bixby. Bixby also supports some fancy features. But Bixby has some limitation, for illustrating, as it supports some limited language due to this, sometimes it fails to provide the direct answer of query.[2]

3) Speech Recognition using Deep Learning and NLP (Natural processing language):



4)

Figure:3 Speech recognition using NLP[11]

Speech recognition software use the various such as NLP (natural language processing) and Deep learning neural network. ”NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way”. NLP is actually a branch of artificial intelligence and particularly deals with the interaction between human and computer. It works on machine learning algorithm and enhances the ability of a computer program to understand human spoken language. It helps the computer to understand and manipulate human language and perform tasks as question answering and language translation.

Deep learning is the subset of Machine learning in A.I that follow the working of human brain in processing data and creating pattern that use in decision making[5].

Related work

T. Al Smadi (May 2015) developed an algorithm based on neural network for speech recognition and its objective is to capture and digitizing the each sound waves, then it converts the each sound waves into the basic language units ,After that it creating words from this unit, and contextually analysing the words to ensure the correct spelling for words .

Advantages of Neural network in speech recognition: Neural networks are the most essential feature of speech recognition. This method is the possibility of parallel processing. Neural network are fed with huge amount of data. Training is given by providing input and educating the network so that it can executed the input and produce the output.

Disadvantages of Neural network in speech recognition: speech recognition are the black box nature of Neural network. Neural network require a lot of data to process as compared to other machine learning algorithms. At least they have millions of labelled sample data. To process this huge amount of data made neural network expensive in terms of size and time complexity. There is no specific rule for determining the structure of a neural network.

ISSN: 2233-7857 IJFGCN

Halageri. A, Bidappa, Arjun, Sarathy and Sultana developed(6 August 2017) an algorithm based on speech recognition using deep learning and its objective is to capture and digitizing the each sound waves, then it converts the each sound waves into the basic language units ,After that it creating words from this unit, and contextually analysing the words to ensure the correct spelling for words . The main objective of this paper is to review the pattern matching abilities of neural networks on speech signal.

Advantages of the existing system are:- Powerful, Self-adjusting, Sophisticated pattern recognition and many more.

Disadvantages of the existing system are, It is not good for device , because it require extra memory to store the data of different individual voice ,as well as it require extra time to execute the task ,so that we can say that it is inefficient in terms of memory and compute time. GMMs are mathematically inefficient for handling the modelling data that lie on or near a nonlinear manifold in the data space. The HMM needs to be trained so that it can execute the user query

ISSN: 2233-7857 IJFGCN

with in time. Apart from that, HMM require set of instruction to execute the datasets.

Advantages of neural network, They can be used to design an input space to any kind of output space. They are simple and, Due to this reason they are commonly used. They are naturally judicial. They are modular in design, so they can be easily attached into larger systems .They have a probabilistic interpretation, so they can be easily combined with statistical techniques like HMMs.

Purwar.K(2015) developed an algorithm based on smart home automation system based on internet of things and speech recognition .The main objectives of this algorithm is to handle the appliances or devices at home. This will provide a better way in automated home as compared to other homes.

Advantages of smart home automation system : This technology is based on internet of things and speech recognition. In this technology we used the command converter which gives the command to the devices. The Raspberry Pi's network and DNS settings, which use the Raspbian operating system to operate and handle the devices. The system is able to use the device at home through speech. There are many benefits of using Raspberry pi as compare to other devices are: It is robust, automated and the capability to run multiple programs. This system is boon for the human being. This technology is fully functional and can be controlled through the wireless system.

Disadvantages of this system is based on internet of things through speech: Some time incompatibility occur between different kind of devices when they are trying to connect each other. Due to this, it increases the complexity between device and also increases the chance of failure of the devices.

Pal Singh .R , Arora. S ,(December 2012)developed an algorithm based on
automatic speech recognition system and objective of this paper is study on various ASR technologies which used in different counties As well the bring the light process which is made for ASR.
1970 Independent approach: Merit points of automatic speech recognition system are the use of the finite network to reduce the computation and determine the closet matching string efficiency. Demerit of automatic speech recognition system are the system recognize speech with vocabulary size of 1011 words with reasonable accuracy.1990 pattern recognition approach: Demerit Several speech recognition error occurred. Merit points: MCC(Minimum classification error) and MMI(Maximum mutual information) both techniques are use reduce the error rate.

Deshmukh .R , Malik Abdullah Alasadi .A,(22 may 2018) developed an algorithm based on automated speech recognition and its objective is study on different speech recognition system and its recent progress also it will describe the characteristics of various database which use in different speech recognition.

Feature extraction:

Feature extraction techniques	**Merit points**	**Demerit points**
MFCC(Mel Frequency Cepstral Coefficients)	It provide better discrimination	Less correlation between coefficients

ISSN: 2233-7857 IJFGCN

DWT(Discrete Wavelet Transform)	Successfully used for noising task. Capacity of compressing a signal without major degradation.	Not based on linear characteristics
WPT(Word Perfect Template)	Same as DWT But WPT show also further detail present in high frequency broad.	Low order coefficient.
LPC(Linear Predictive Codes)	LPC is easy to implement and mathematical precise	Static feature Extraction

5) Comparison between some popular Existing Speech Recognition systems

Table 1: Comparison between some popular Existing Speech Recognition systems [3]

Differentiable Factors	**Apple Siri**	**Google Assistant**	**Microsoft CORTANA**	**Amazon Alexa**	**Samsung Bixby**
Release date	2011	2016	2014	2014	2017
Device compatibility	Apple devices	Android devices	Microsoft devices	Alexa and echo	Samsung phones
Unsupported apps	Google mail services	Support all third party apps	Youtube	Youtube	Support all third party apps
Type of connection Required	WIFI and cellular data	WIFI and cellular data	WIFI and cellular data	Only WIFI	WIFI and cellular data
Ways to interact	Manual and voice commands	Manual and voice commands	Manual and voice commands	Manual (in silent echo) voice	Manual and voice commands

ISSN: 2233-7857 IJFGCN

				commands	
Algorithms	Dynamic tim e wrapping, discretization algorithm.	PLP features, Viterbi search, Deep Neural Networks	natural langua ge processing (NLP) , data is sent to Microsoft's servers to be analyzed	natural langua ge processing (NLP)	discriminative training, WF ST framework

ISSN: 2233-7857 IJFGCN

Accuracy according to search results	74.6	88	63	72.5	78
Languages supported	21	40	8	7	7
Major drawbacks	Only on when hold the home button.	Maximum battery usage and slow down the device	Vulnerability found hit the listening button again and again	Take several days to install or schedule a new version of software	Server less extensibility , NLP platform

Different speech recognition system are compared and descriptions are mentioned in the above table. After studying these different application systems some drawbacks are found in these existing systems such as most of the time, some recognition system are not be able to work with every languages ,they only support some limited languages. Apart from that, some individual has a habit of speak very fast with a strong accent, in that cases some systems are not be able to handle the user task .Even though, some recognition system create huge background noise during user request .Due to this, they are not be able to attain the user request on that moment that's why individual has to speak again and again until system do not respond the user query.

ANALYTICAL STUDY FOR DIFFERENT TRAINING DATASET:

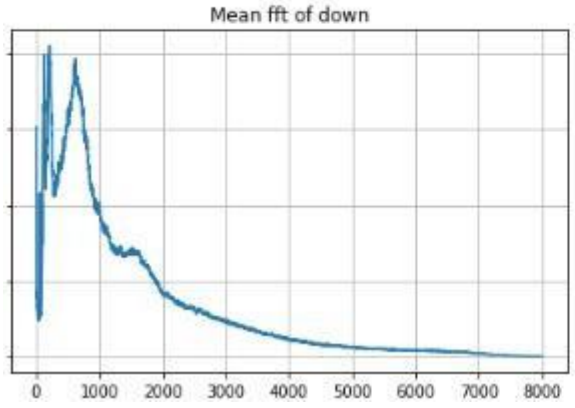
Firstly, we compared different research papers of the speech recognition system. On behalf of this paper we find out different training datasets, then we executed these datasets on the IDLE.
Different training datasets are available below in form of screenshot which we executed on the IDLE and JUPITER notebook platform. Here we find out different dataset produce the different scenarios of speech. This situation can be highlighted by an example, many individuals provide the same input in form of speech, with the help of these training dataset we executed the input and it produce the several outputs rather than to produce the single output and the reason behind for each individual belongs to a different country each individual has own accents. Due to this, speech recognition system is not able to determine the voice (input) of each individual.[17]

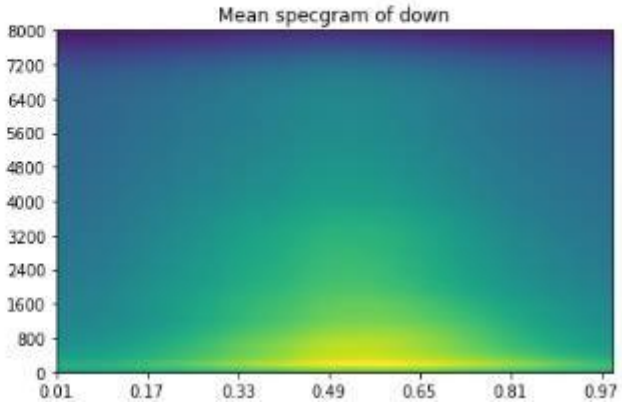
ISSN: 2233-7857 IJFGCN

ATTRIBUTES OF DIFFERENT TRAINING DATASETS:

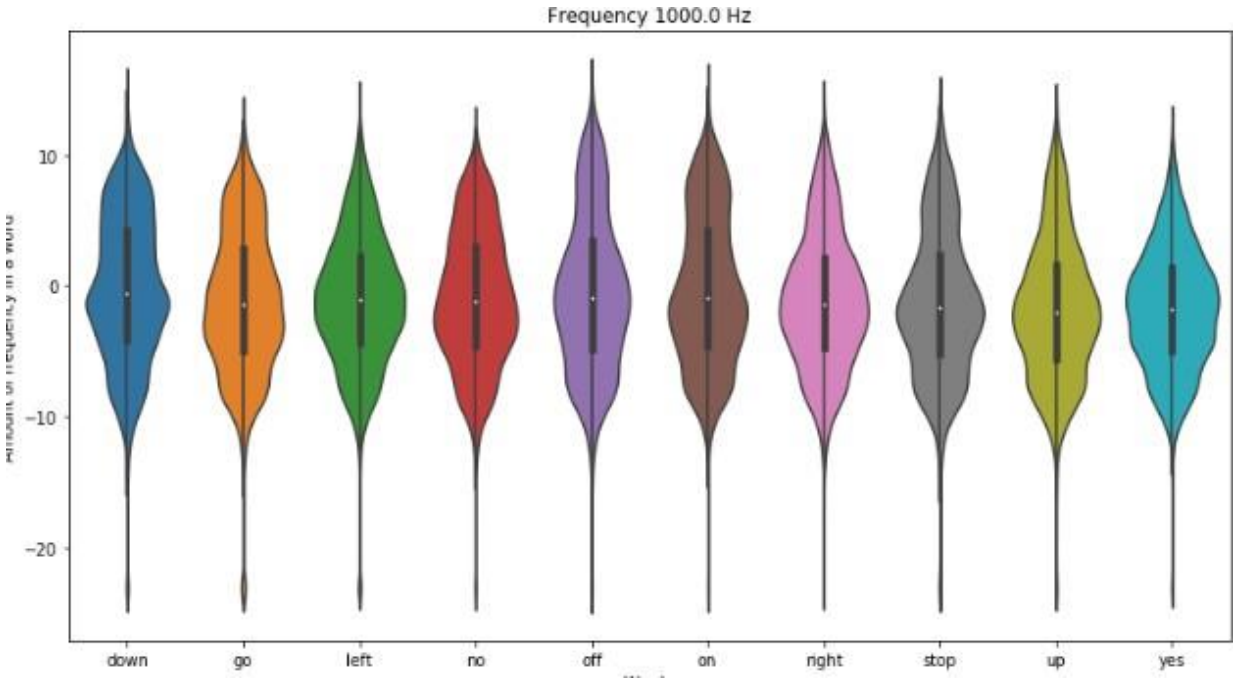
Dataset	Size(in MB)	Wavelength(in meter)	Frequency(in Hz)	Channel	Bitrate(in kbps)	Sampling rate(in kHz)
Afrikaan	1.74	3.811	90	1	705	44.1
Arabic	3.82	2.144	160	1	705	44.1
Bambara	2.99	2.318	148	1	705	44.1
Catalan	2.34	2.579	133	1	705	44.1
Dutch	2.97	2.257	152	1	705	44.1
English	1.83	3.333	103	1	705	44.1
Hindi	2.50	2.433	141	1	705	44.1
Farsi	2.30	2.701	127	1	705	44.1
German	2.21	3.090	111	1	705	44.1
Italian	2.52	2.199	156	1	705	44.1

```
['down', 'go', 'left', 'no', 'off', 'on', 'right', 'stop', 'up', 'yes']
```





ISSN: 2233-7857 IJFGCN



DIFFERENT OUTPUTS OF DIFFERENT DATASETS:

TRAINING DATA SET1 OUTPUT:-

```
============== RESTART: C:\Users\hari om\Documents\prctice.py ==============
audio file contains please call Stella to bring the scenes with her from the store s
rk unit is mole plastic snake and frog for the kids scoop the scenes in 2-3 redbacks
```

TRAINING DATASET2 OUTPUT:

```
============== RESTART: C:\Users\hari om\Documents\prctice.py ==============
audio file contains please call Stella ask her to bring this things with her from
also need small plastic snake and the big frock for kids she can cook this things
```

TRAINING DATASET3 OUTPUT:

```
============== RESTART: C:\Users\hari om\Documents\prctice.py ==============
audio file contains pic of Dalla to bring these things with her from the store
ake in a big doll frock with kids scooties things into 3 red bags Navami Havan
```

TRAINING DATASET4 OUTPUT:

```
============== RESTART: C:\Users\hari om\Documents\prctice.py ==============
audio file contains please call Stella ask her to bring these things with her
ts for her brother Bob we also need a small plastic snake and big toy for the
```

TRAINING DATASET5 OUTPUT:

```
============== RESTART: C:\Users\hari om\Documents\prctice.py ============
audio file contains please call Stella uska to bring this thing this things
nake for a brother power we also need a small plastic snake and a big toy f
```


ISSN: 2233-7857 IJFGCN

RESULT OF DIFFERENT TRAINING DATASET:

From the above screenshots, it is concluded that limited training data which we use in our code is not able to determine the exact output what user expect from the system. Each training data sets take the same inputs but it produces different output. We provided several individual voices but it is not able to distinguish each individual voice because each individual has own accents. In future our objective is to remove these drawbacks by developing a new application which is able to communicate easily to us, it doesn't require a lot of human effort to make that machine-understandable or it can be easily controlled by our voice and doesn't require any buttons to be pressed. The application should be light, handy, and easy to use means it doesn't use a lot of battery, storage, and memory which basically affects the

performance of the device. The main motive of the application is to perform every single function on the device that our fingers are able to do.

6) Conclusion and Future scope

In this paper, we give a brief introduction to speech recognition, how does it works and some popular systems that works on speech recognition algorithms. A comparative analysis of various speech recognition systems are also discussed in this paper. This paper also helps in having an idea of current best system on speech recognition by analyzing their performance according to the different differentiable factors. The discussion may be used to improve or develop a new or modified system which is more accurate and perform better than these current running systems.

References:

1) Gonfalonieri,A.,2018,Toward Data science <https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3>
2) Hall, C& Tillman, M, 13 march 2020,Pocket-Lint <https://www.pocket-lint.com/phones/news/samsung/140128-what-is-bixby-samsungs-assistant-explained-and-how-to-use-it>
3) August 9,2019,Reachbyte, <https://reachbyte.com/siri-cortana-alexa-google-assistant-bixby>
4)Goel, A., February 2,2018,Artifical Intelligence,<https://magoosh.com/data-science/siri-work-science-behind-siri>
5) Le,J.,Sep20,2019,Deep learning-based Automatic speech recognition,<https://heartbeat.fritz.ai/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380>
6) Prospero, M. & Kozuch, K., February 24,2020,The best Google assistant skills,<https://www.tomsguide.com/round-up/best-google-assistant-features>
7) Hui ,J., Sep 16,2019,Language model,<https://medium.com/@jonathan_hui/speech-recognition-acoustic-lexicon-language-model-aacac0462639>
8) Sarma , M. & Sarma,K.,12 sep,2015,'Acoustic Modelling of speech signal using Artificial Neuron Network :A review of techniques and current Trends,DOI:10.4018/978-1-4666-8493-5.ch012
9)Rudnicky,1July,2010,language Model,<http://www.voxforge.org/home/docs/faq/faq/what-is-a-language-model>
10) Saba,G.,22July,2018,Speech recognition Python,<https://www.simplifiedpython.net/speech-recognition-python>
11) NDZ,28 April,2017,Machine Learning Vs Artificial Intelligencehttps://ndimensionz.com/kb/theoretical-explanation-of-how-siri-works
12) T. Al Smadi et al, Journal of Signal and Information Processing, 2015, 6, 66-72 Published Online May 2015 in SciRes. http://www.scirp.org/journal/jsip http://dx.doi.org/10.4236/jsip.2015.62006
13)Halageri.A et al, / (IJCSIT) International Journal of Computer Science and Information

ISSN: 2233-7857 IJFGCN

Technologies, Vol. 6 (3) , 2015, 3206-3209
14) Purwar.K, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.6, August 2017
15) R. pal Singh ,S. Arora , December 2012, Automatic speech recognition system :A review,vol:60
16) R.Deshmukh ,A.malik Abdullah Alasadi,22 may 2018,Atomated speech recognition: A review.
17)R.Tatman,2017,machine training dataset< https://www.kaggle.com/rtatman/speech-accent-archive>

ISSN: 2233-7857 IJFGCN