Knowledge Consortium of Gujarat
Department of Higher Education - Government of Gujarat
Journal of Science - ISSN : 2320-0006

Continuous Issue - 12 | June – July 2017

Implementation of Automatic Speech Recognition tool using C#.Net

Abstract

In Current scenario, human beings always look forward on something new, something change. Automatic speech recognition is the one of the topic which is in underdevelopments. Google, IBM such companies are also working on it and wants to take this system to advance level. Keeping in mind the current rising trend of automatic speech recognition I have propose the following software. A software which can communicate with user and perform them task depending upon the voice command like a virtual assistant. That can able to recognize human voice and performing the task, it can also type that you are speaking as well as it can able to speak whatever you have typed. It can also give response of your answer.

Keywords: *Python, Java, Automatic Speech Recognition, CMU Sphinx, speaker- independent, Pattern recognition approach, Directed Dialogue Conversion.*

I. Introduction

Automatic speech recognition is the one of most trending concepts. We can say it is our bright future, because nowadays people want the task to become as easy as possible as less as well as with fun. The salient feature of this software is that the commands are“a part of our day to day conversation as well as few different commands”.

How does it feel whenwe are talking with our computer and laptop? It's a fun know. I guess in nearer decades most of task will be best one voice commands whether it's home, school, college & universities, company or industries all will be based on voice commands, in which you just have to order your task to your computer and it will complete task for you easily. Even Google is going to making a new project named Google Home Assistant which can help you to do make task easier and with more new facilities. Same as online shopping website Amazon lunched Amazon ECO, an assistant that can help your different kinds, like it will play music for you, switch on light or fan and many more thing. It can also help you to buy anything online from Amazon. All this thing to be done by just voice commands. In fact, IBM is also working in the project for making a virtual assistant with Artificial Intelligent and Natural Language Processing(NLP) concepts. It may be fun to know that Mark Zuckerberg, the CEO of Facebook also wants to build a personal virtual assistant for himself. Apple's Siri, Microsoft's cortnata and Google's OK Google are some available popular smart assistants. But unfortunately, none of them are working without internet. They are worth less without internet.

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. Since the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis[1].

Actually, Speech recognition can be covered in Electronics & Instrumental department but I'm from Computer Science so I'll focus on programming method rather than dealing with frequencies and all. Speech recognition can be possible in Matlabalso and high-level programming language also like JAVA, PYTHON and many more. It's much efficient to make speech recognition system in high level language. I'm suggesting you choose Python rather than java because Python is much easier to implement and reduce the coding. The most popular toolkit for "Acoustic Model" training is HTK (Hidden Markov

Model ToolKit).[7]. The speech recognition engine is in two versions, written in C and in Java for acoustic model training tools. These acoustic model training tools are conformable to many programming new language.[8]

Proposed model:

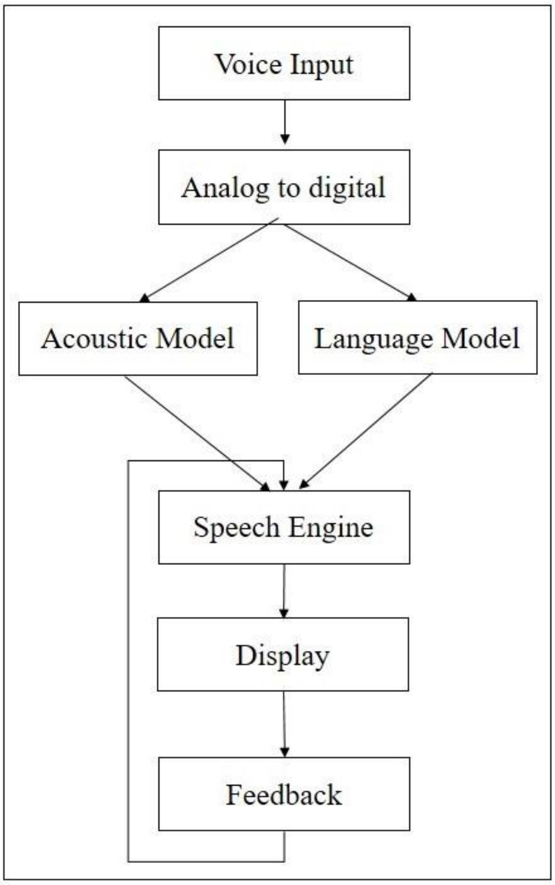
In easier language it can be defined as, voice recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format [4]. Following is basic structure of Automatic voice recognition system:

Acoustic models:

An acoustic model is created by taking audio recording of speech and their text transcriptions and using software to create statistical representation of the sounds that make up each word.

Language models:

Language model is used in many natural processing applications such as speech recognition which tries to capture the properties of a language and to predict the next word in a speech sequence.



Now, let's we are going forward types of speech recognition,

➢ **Type of voice recognition:**

- Speaker - Dependent:

- It works by learning the unique characteristics of single person's voice in a way similar to voice recognition.

- So that new users must first "train" the software by speaking, so that system can analyze how the person talks.

- This often means user have to read a few pages of text to the computer before they use.

- Speaker -Independent:

- It is designed to recognize anyone's voiceso no training is involved.

- So it directly means that user don't have to read no of pages of text before using the system.
- The downside is that speaker – independent software is generally less accurate than dependent software.

Speech recognition engines that are speaker independent generally deal with this fact by limiting the grammars they use. By using a smaller list of recognized words, the speech engine is more likely to correctly recognize what a speaker said.

➢ **Different processes involved:**

- Digitization :
- Analog to digital conversion
- Sampling and quantization

- Signal processing :
- Separating speech from back ground noise.

- Phonetics :
- Variability in human speech.

- Phonology :
- Recognizing individual sound distinctions(Similar phonemes)
- It is systematic use of sound to encode meaning in any spoken human language.

➢ **Approaches to speech recognition**

There are three types of approaches to ASR. They are[5]:

- Acoustic phonetic approach
- Pattern Recognition approach
- Artificial intelligence approach.

❖ Acoustic Phonetic Approach

- Acoustic phonetic approach is also known as rule-basedapproach. This approach uses knowledge of phonetics &linguistics to guide search process. There are usually somerules which are defined expressing everything or anything thatmight help to decode based in "blackboard" architecture i.e. ateach decision point it lays out the possibilities and apply rulesto determine which sequences are permitted. It has poorperformance due to difficulty to express rules, to improve thesystem. This approach identifies individual phonemes, words,sentence structure and/or meaning.

❖ Pattern Recognition Approach:

- This method has two steps i.e. training of speech patternsand recognition of pattern by way of pattern comparison. Inthe parameter measurement phase (i.e filter bank), a sequence of measurements is made on the input signal todefine the "test pattern". The unknown test pattern is thencompared with each sound reference pattern and a measure ofsimilarity between the test pattern & reference pattern bestmatches the unknown test pattern based on the similarityscores from the pattern classification phase (dynamic timewarping).

❖ Template Matching Approach:

- Test pattern T, and reference pattern {R1,…,Rv} are represented by sequences of feature measurements. Patternsimilarity is determined by aligning test pattern T with reference pattern Rv with distortion D(T,Rv). Decision rulechooses reference pattern R* with smallest alignment distortion D(T,R*).R*= argmin D(T,Rv)Dynamic Time Warping (DTW) is used to compute the bestpossible alignment v between T and Rv and the associated distortion D(T, Rv).\

❖ Stochastic based approach:

- It can be seen as extension of template based approach, usingsome powerful and statistical tools and sometimes seen asanti-linguistic approach. It collects a large corpus oftranscribed speech recording and train the computer to learnthe correspondences. At run time, statistical processes areapplied to search for all the possible solutions & pick the bestone.

❖ Artificial Intelligence Recognition Approach:

- This approach is a combination of the acoustic phoneticapproach & the pattern recognition approach. In the AI, anexpert system implemented by neural networks is used toclassify sounds. The basic idea is to compile and incorporateknowledge from a variety of knowledge sources with theproblem at hand.

Of the two types we selected speaker – independent voice recognition.

Now further, there are two more type of it which are following [2]

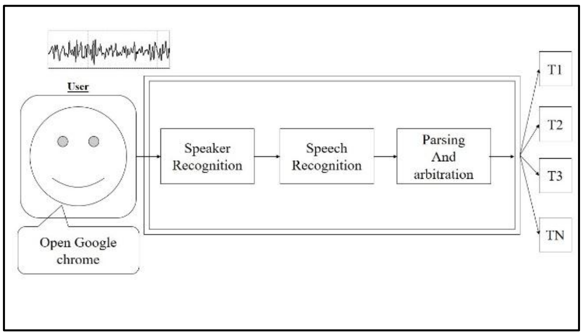
- **Directed Dialogue conversations**

- These are the much simpler version of ASR at work and consist of machine interfaces what tell you verbally to respond with a specific word from a limited list of choices, thus forming their response to your narrowly defined request.
- Automated telephone banking and other customer service interfaces commonly use directed dialogue ASR software.

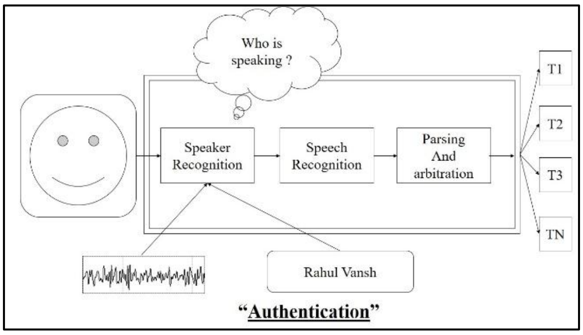
- **Natural Language Conversations**

- These are the much more sophisticated variants of ASR and instead of heavily limited menus of words you may use, they try to simulate real conversation by allowing you to use an open-ended chat format with them.
- The Siri interface on the iPhone is a highly advanced example of these systems.

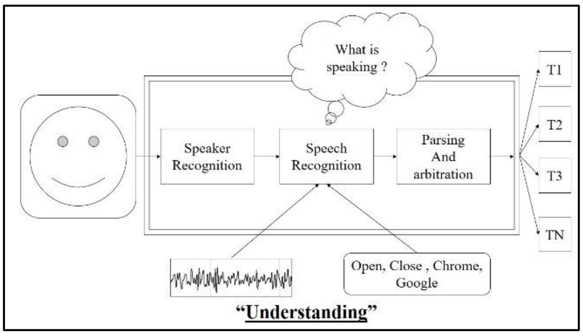
➢ The basic sequence of events that makes any Automatic Speech Recognition software, regardless of itsSophistication, pick up and break down your words for analysis and response, goes as follows:



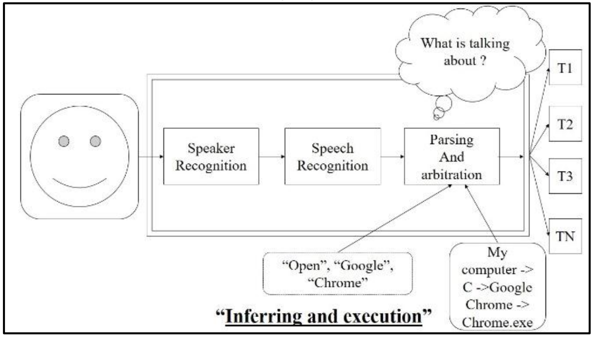
(i)



(ii)



(iii)


What is talking about ?
Speaker Recognition
Speech Recognition
Parsing And arbitration
T1
T2
T3
TN
"Open", "Google", "Chrome"
My computer -> C ->Google Chrome -> Chrome.exe
"Inferring and execution"

(iv)

- You speak to the software via an audio feed
- The device you're speaking to creates a wave file of your words
- The wave file is filtered by removing background noise and normalizing volume
- The resulting filtered wave form is then broken down into what are called phonemes. (Phonemes arethe basic building block sounds of language and words. English has 44 of them, consisting of sound blocks such as "wh", "th", "ka" and "t".
- Each phoneme is like a chain link and by analyzing them in sequence, starting from the first phoneme, the ASR software uses statistical probability analysis to deduce whole words and thenfrom there, complete sentences.
- Your ASR, now having "understood" your words, can respond to you in a meaningful way. Here, T1,T2…Tn is no of task.

➢ Available library tools[3]:

- Speech to text engines:

- **Pocketsphinx**is a open-source speech decoder by the CMU Sphinx project. It's fast and designed to work well on embedded systems (like the Raspberry Pi).On the other hand, recognition will be performed offline, i.e. you don't need an active internet connection to use it. It's the right thing to use if you're cautious with your personal data.
- **Google STT**is the speech-to-text system by Google. If you have an Android smartphone, you might already be familiar with it, because it's basically the sameenginethat performs recognition if you say OK, Google. It can only transcribe a limited amount of speech a day and needs an active internet connection.
- **AT&T** STT is a speech decoder by the telecommunications company AT&T. It's online and thus needs an active internet connection.
- **Julius**is a high-performance open source speech recognition engine. It does not need an active internet connection.

- Text to speech engines:

- **eSpeak**is a compact open-source speech synthesizer for many platforms. Speech synthesis is done offline, but most voices can sound very "robotic".
- **Festival** uses the Festival Speech Synthesis System, an open source lib. Like eSpeak, also synthesizes speech offline.

- **Flite** uses CMU Flite (festival-lite), a lightweight and fast synthesis engine that was primarily designed for small embedded machines. It synthesizes speech offline, so no internet connection is required.
- **SVOX Pico TTS** was the Text-to-Speech engine used in Android 1.6 "Donut". It's an open-source application and also works offline. The quality is rather good compared to eSpeak and Festival.
- **Google TTS** uses the same Text-to-Speech API which is also used by newerAndroid devices. The Synthesis itself is done on Google's servers, sothat youneed an active internet connection and also can't expect a lot of privacy if you usethis.
- **Ivona TTS** uses Amazon's Ivona Speech Cloud service, which is used in the KindleFire. Speech synthesis is done online, so an active internet connection andAmazon has access to everything Jasper says to you.
- **MaryTTS**is an open-source TTS system written in Java. You do not need internet access.
- **Mac OS X TTS** does only work if you're running Jasper on a Mac. It then uses thesay command in MacOS to synthesize speech.

➢ **Purpose:**

- To make a system which works on voice commands. A software which takes commands from user and does that task which we want.
- It's like a virtual assistant which does task like a personal assistant. It can be also useful in first type because here we don't have to type anything we just have to speak and it gives 2X faster speed then typing.

➢ **Props:**

- Computer/Laptop
- Good quality headphone
- Microphone
- Raspberry Pi

- In below, proposed model I'm going to choose Speaker independent-> Pattern Recognition Approach -> Directed Dialogue Conversion respectively.
- Lib tools such as CMU Sphinx, Google STT for STT and eSpeak or MaryTTS TTS are some good libraries. I prefer to choose specially CMU Sphinx because by using this library we can able to use it without internet [6] and it's mostimportantthing that I like to implement. Because most of software doesn't work without internet like Cortana, Siri.
- Programming language: Python

➢ **Working ideas:**

- My software name is Stark and here I discuss some commands. It is built on speaker – independent and Directed Dialogue conversationsmethod. I'm putting "Stark" before the commands, it's because after telling "Stark" it will start to recognize your commands. Before, using you have to make your profile in which you have to give information like name, age, gender etc. so that Stark can reply such as "Okay sir"/ "Okay madam"

- It can work on different task very efficiently; one of very common task is to convert voice to text in Microsoft word, notepad, word pad etc.

- It can also tell you time as well as weather according prevailing conditions.

- Now, if you don't like reading long articles then don't worry, we just have to select text and give command like "Stark, speak" and it will speak out to you.

- Ifyou tell to Stark to introduce itself by telling "Stark, introduce yourself" then it will give its own introduction by telling its name and its usefulness.

- You can also greet Stark such as, good morning, good evening, good night. But if you greet wrongly then it will correct you. Like if you tell "Stark, Good evening" at 6:00 AM then it'll reply "Sorry sir, Its<current time>so you must say Good morning".

Here, I'm going to show you basic working strategy:

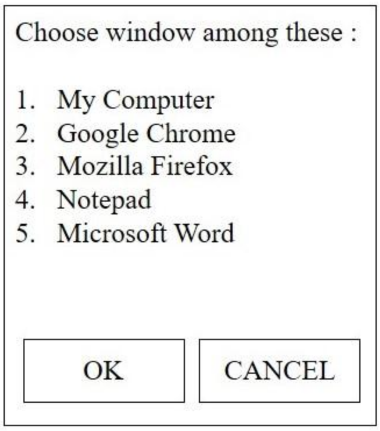
```
// ch= spoken string
if(<time>> 5:00 &&<time><12:00)
{
    if(”ch” == “Good morning”)
    {
    reply(“Good morning”);
    }
    else
    {
reply(“Sorry, It’s <current time>    So you must say Good Morning”);
    }

}
```

Such as,

- For 12:00 to 17:00 reply” Good Afternoon”.
- For 17:00 to 21:00 reply “Good Evening”
- For 21:00 to 00:00 reply “Good Night”

Above algorithm will not run but just used to give you basic idea of what I want to convey actually.

- You can switch OFF/ON by speech as well as by few clicks on icon. Now, suppose it you want toggle to sleep mode then we have to tell “Stark, Sleep” then it’ll reply “Okay sir, Bye”. Such as if you want to on again it then you have to speak “Stark, Wake up” then it’ll reply “Stark is active” - “I’m on”.
- Now if you want know something like, who is Steve Jobs? Then it’ll open your browser. i.e. Google chrome, Mozilla Firefox then give reply “Sir, its somewhat that you want to you want to find” for this you have to tell “Stark, Search, who is the Steve Jobs?”
- During speaking, in the sense of typing in Microsoft Word or Notepad if the software will get confused like between ‘weather’ and ‘whether’ then it will ask you to choose among both.
- While finding or asking anything ifit can’t reply without internet then software will speak “Sorry sir, I’m not able to reply please connect to internet”.
- You can also show command list by speaking “Stark, Show command list”. The software also facilitates you to modify or add more command. For that you have to teach your software by entering speech and giving appropriate directory to perform any particular task. You can also type text that you want to get in reply during this command but it’ll be optional.
- While doing multiple task, if you want to toggle to other window then you have to speak “Stark, Show me the running windows”. Now you have to choose among which window you want to work. Likewise below ,



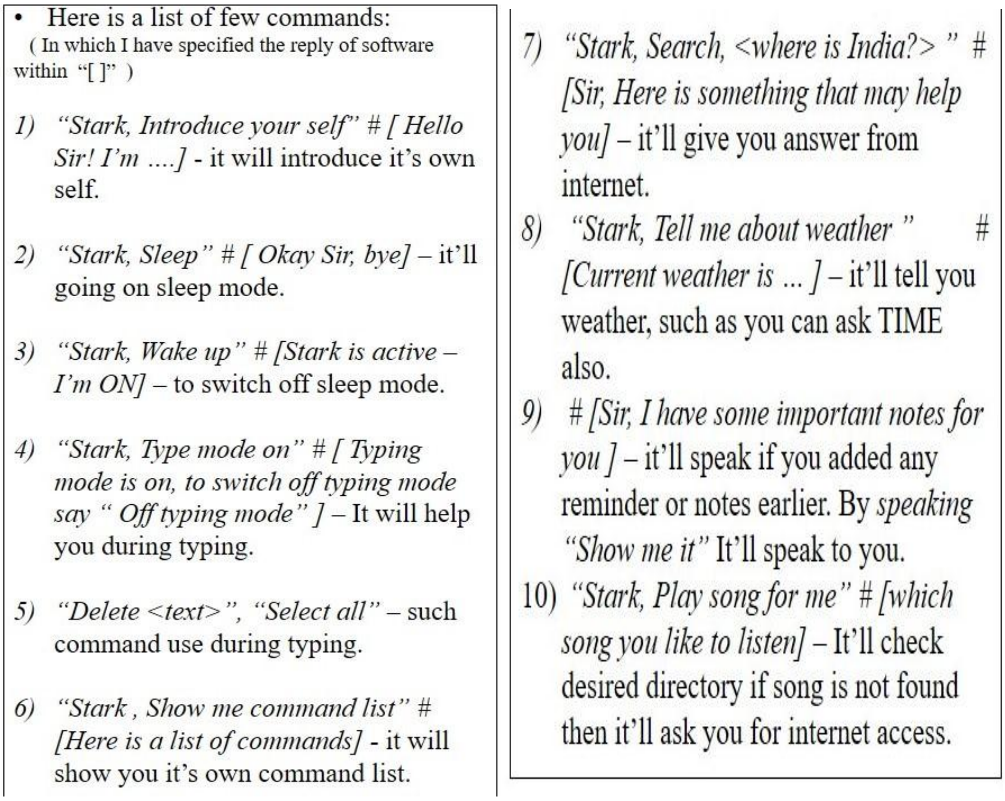
Choose window among these :

1. My Computer
2. Google Chrome
3. Mozilla Firefox
4. Notepad
5. Microsoft Word

You can enter your choice by speaking 1, 2 or OK, CANCEL
Which you want to choose.

Command list :

- This list I made by my own self you can also make your own different list as you like build.



- Here is a list of few commands:

(In which I have specified the reply of software within "[ ]")

1) *"Stark, Introduce your self" # [Hello Sir! I'm ....]* - it will introduce it's own self.
2) *"Stark, Sleep" # [Okay Sir, bye]* – it'll going on sleep mode.
3) *"Stark, Wake up" # [Stark is active – I'm ON]* – to switch off sleep mode.
4) *"Stark, Type mode on" # [Typing mode is on, to switch off typing mode say " Off typing mode"]* – It will help you during typing.
5) *"Delete <text>", "Select all"* – such command use during typing.
6) *"Stark , Show me command list" # [Here is a list of commands]* - it will show you it's own command list.
7) *"Stark, Search, <where is India?> " # [Sir, Here is something that may help you]* – it'll give you answer from internet.
8) *"Stark, Tell me about weather " # [Current weather is ...]* – it'll tell you weather, such as you can ask TIME also.
9) *# [Sir, I have some important notes for you]* – it'll speak if you added any reminder or notes earlier. By *speaking "Show me it"* It'll speak to you.
10) *"Stark, Play song for me" # [which song you like to listen]* – It'll check desired directory if song is not found then it'll ask you for internet access.

- In above command list, you may observed, that I used "Stark" at the beginning of every commands because after speaking "Stark" software starts to recognize your commands.

Characteristics and the possibility:

- It is possible to distinguish the following most importantcharacteristics and the possibility of using smart terminal devices [4]:
- Voice search – search information stored on the mobile terminal device
- Voice calling – voice dialing and calling the contact or number
- Search/Internet – finding the required information by using voice commands within a web browser of mobile terminal device
- Voice-to-Text convert – voice commands and messages sent by the speaker (user of mobile terminal device) are converted into text form and can be used to send messages in the form of text messages, e-mail messages etc.
- Voice reproduction – ability for easier reception and understanding of some of the messages (SMS, e-mail,
instant messaging) in a way that message received in text format is automatically reproduced by using voice
- Messages finding – potential use of voice when searching information in a way that certain messages (SMS
messages, reminders, calendar information, e-mail messages) filtered and used in accordance with the finding keywords
- Calendar/Reminders – add / delete / edit memos and obligations related to the calendar of the mobile terminal device by using voice
- Add/change notes – voice option of adding, deleting and editing notes of the user of mobile terminal device
- Weather report – weather forecast and weather monitoring application managed by voice
- Multimedia access – smart mobile terminal devices are multifunctional devices that provide numerous multimedia data (video files, music files, photo gallery, etc.) which can be easily accessed using the human voice

- **Limitation:**

- Noise robustness
- Speaker/Accent channel
- Language model mismatches
- Web is multi-lingual

❖ Here, I'm going to give you references of few websites which may help you in implementation.

- http://usabilitygeek.com/automatic-speech-recognition-asr-software-an-introduction/
- http://jasperproject.github.io/documentation/configuration/
- https://wolfpaulus.com/journal/embedded/raspberrypi2-sr/
- http://cmusphinx.sourceforge.net/wiki/faq
- https://oscarliang.com/raspberry-pi-voice-recognition-works-like-siri/
- https://diyhacking.com/best-voice-recognition-software-for-raspberry-pi/

Conclusion

- In this paper, I have made an attempt to brief you about the multifaceted speech recognition technique whose commands are "a part of our day to day conversation as well as few different commands". Speech recognition software though difficult to design is one of the most reliable and efficient tools in recent times.

References

I. Automatic Speech Recognition â€“ A Brief History of the Technology Development (2004) by B. H. Juang, Lawrence R. Rabiner edited by Georgia Institute of Technology, Atlanta & Rutgers University and the University of California, Santa Barbara

II. http://usabilitygeek.com/automatic- speech-recognition-asr-software-an-introduction/

III. http://jasperproject.github.io/documentation/configuration/

IV. 4-- find correct from science direct

V. PreetiSaini ,Parneet Kaur. "Automatic Speech Recognition: A Review". International Journal of Engineering Trends and Technology (IJETT). V4(2):132-136 Feb 2013. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group

VI. http://cmusphinx.sourceforge.net/wiki/faq

VII. P. Nguyen, "Techware: Speech recognition software and resources on the web [Best of the Web]," in IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 102-105, May 2009. doi: 10.1109/MSP.2009.932160

VIII. P. K. Sahu and D. S. Ganesh, "A study on automatic speech recognition toolkits," 2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE), Bhubaneswar, 2015, pp. 365-368.doi: 10.1109/ICMOCE.2015.7489768

[4--SinisaHusnjak*, DraganPerakovic, Ivan Jovovic. Possibilities of using Speech Recognition Systems of Smart Terminal Devices in Traffic Environment.

****

Ms. Ripal Ranpara
Assistant Professor
Department of Computer Science & IT
Shree M. & N. Virani Science College (Autonomous)
Rajkot