AI Based Voice Assistant Using Speech Recognition

Anup Bhange
Assistant Professor
Dept. of Computer Technology, KDK College of Engineering, Nandanwan, Nagpur

Deepak Shende
Dept. of Computer Technology, KDK College of Engineering, Nandanwan, Nagpur

Ria Umahiya
Dept. of Computer Technology, KDK College of Engineering, Nandanwan, Nagpur

Monika Raghorte
Dept. of Computer Technology, KDK College of Engineering, Nandanwan, Nagpur

Aishwarya Bhisikar
Dept. of Computer Technology, KDK College of Engineering, Nandanwan, Nagpur

Abstract— **Artificial intelligence technologies are beginning to be actively used in human life, this is facilitated by the appearance and wide dissemination of the Internet of Things (IOT). Autonomous devices are becoming smarter in their way to interact with both a human and themselves. New capacities lead to creation of various systems for integration of smart things into Social Networks of the Internet of Things. One of the relevant trends in artificial intelligence is the technology of recognizing the natural language of a human. New insights in this topic can lead to new means of natural human-machine interaction, in which the machine would learn how to understand human's language, adjusting and interacting in it. One of such tools is voice assistant, which can be integrated into many other intelligent systems.**

In this paper, the principles of the functioning of voice assistants are described, its main shortcomings and limitations are given. The method of creating a local voice assistant without using cloud services is described, which allows to significantly expand the applicability of such devices in the future

Index Terms: Voice assistant, Speech Recognition, Low cost, Internet, Speech Synthesis, Visually Challenged.

I. INTRODUCTION

Today the development of artificial intelligence (AI) systems that are able to organize a natural human-machine interaction (through voice, communication, gestures, facial expressions, etc.) are gaining in popularity. One of the most studied and popular was the direction of interaction, based on the understanding of the machine by the machine of the natural human language. It is no longer a human learns to communicate with a machine, but a machine learns to communicate with a human, exploring his actions, habits, behavior and trying to become his personalized assistant.

The work on creating and improving such personalized assistants has been going on for a long time. These systems are constantly improving and improving, go beyond personal computers and have already firmly established themselves in various mobile devices and gadgets. One of the most popular voice assistants are Siri, from Apple, Amazon Echo, which responds to the name of Alex from Amazon, Cortana from

Microsoft, Google Assistant from Google, and the recently appeared intelligent assistant under the name "AIVA".

Section I, II presents a brief introduction to the architecture and construction of voice assistants. Section III provides proposed plan of work. Section IV provides methodology of the work of a voice assistant AIVA. Section V describes the test results of the voice assistant. Section VI and VII describes the conclusion and future scope of an assistant using various artificial intelligent algorithms, and gives a comparative evaluation of the learning ability of algorithms. The main goal of this work is to build a local voice assistant that does the work of human and the daily task that a human needed to do in daily life.

AIVA (2018) aimed at developing a voice-controlled personal assistant which is doing many things such as to search the Internet. It has some new features like posting comments on the social media websites such as Facebook, Twitter, etc. By just few simple commands. You can also know the weather around you and can get the climate conditions in your region. It can open and launch web-applications and the local storage of the user computer.

II.RELATED WORK

Each company-developer of the intelligent assistant applies his own specific methods and approaches for development, which in turn affects the final product. One assistant can synthesize speech more qualitatively, another can more accurately and without additional explanations and corrections perform tasks, others are able to perform a narrower range of tasks, but most accurately and as the user wants. Obviously, there is no universal assistant who would perform all tasks equally well. The set of characteristics that an assistant has depends entirely on which area the developer has paid more attention. Since all systems are based on machine learning methods and use for their creation huge amounts of data collected from various sources and then trained on them, an important role is played by the source of this data, be it search systems, various information sources or social networks. The amount of information from different sources determines the nature of the assistant, which can result as a result. Despite the different approaches to learning, different algorithms and techniques, the principle of building such systems remains approximately the same. Figure 1 shows the technologies that are

used to create intelligent systems of interaction with a human by his natural language. The main technologies are voice activation, automatic speech recognition, Teach-To-Speech, voice biometrics, dialog manager, natural language understanding and named entity recognition.

VOICE TECHNOLOGY	BRAIN TECHNOLOGY
Voice Activation	Voice Biometrics
Automatic Speech Recognition (ASR)	Dialog Management
(Teach-To-Speech (TTS)	Natural Language Understanding (NLU)
	Named Entity Recognition NER)

Fig.1. Technologies for constructing intelligent systems of interaction with a human by natural language.

III. PROPOSED PLAN OF WORK

The work started with analyzing the audio commands given by the user through microphone. This can be anything like getting any information, operating computer's internal files, etc. This is an empirical qualitative study, based on reading above mentioned literature and testing their examples. Tests are made by programming according to books and online resources, with the explicit goal to find best practices and a more advanced understanding of Voice Assistant.

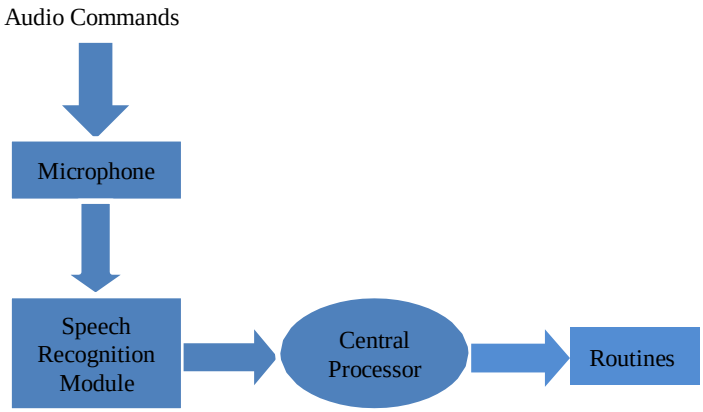


Fig.2 Basic Workflow

Fig.2 shows the workflow of the basic process of the voice assistant. . Speech recognition is used to convert the speech input to text. This text is then fed to the central processor which determines the nature of the command and calls the relevant script for execution.

But, the complexities don't stop there. Even with hundreds of hours of input, other factors can play a huge role in whether or not the software can understand you. Background noise can easily throw a speech recognition device off track. This is because it does not inherently have the ability to distinguish the ambient sounds it "hears" of a dog barking or a helicopter flying overhead, from your voice. Engineers have to program that ability into the device; they conduct data collection of these ambient sounds and "tell" the device to filter them out. Another factor is the way humans naturally shift the pitch of their voice to accommodate for noisy environments; speech recognition systems can be sensitive to these pitch changes.

IV. METHODOLOGY

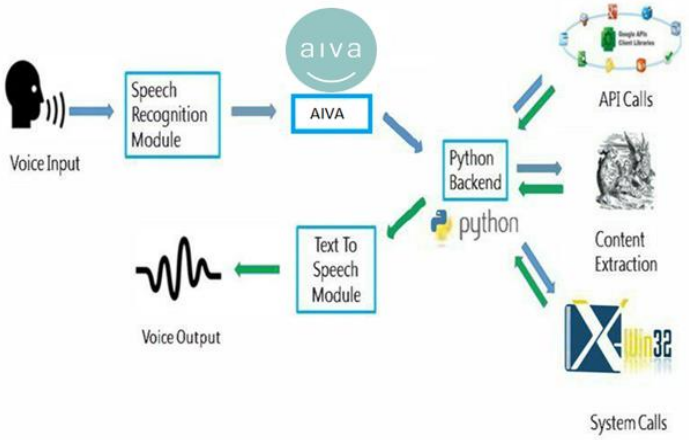


Fig.3. Detailed Workflow

A. *Speech Recognition*

The system uses Google's online speech recognition system for converting speech input to text. The speech input Users can obtain texts from the special corpora organized on the computer network server at the information center from the microphone is temporarily stored in the system which is then sent to Google cloud for speech recognition. The equivalent text is then received and fed to the central processor.

B. *Python Backend*

The python backend get the output from the speech recognition module and then identifies whether the command or the speech output is an API Call, Context Extraction, and System Call. The output is then send back to the python backend to give the required output to the user.

C. *API Calls*

API stands for Application Programming Interface. An API is a software intermediary that allows two applications to talk to each other. In other words, an API is the messenger that delivers your request to the provider that you're requesting it from and then delivers the response back to you.

D. *Context Extraction*

Context extraction (CE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as context extraction TEST RESULTS.

E. *System Calls*

In computing, a system call is the programmatic way in which a computer program requests a service from the kernel of the operating system it is executed on. This may include hardware-related services (for example, accessing a hard disk drive), creation and execution of new processes, and communication with integral kernel services such as process scheduling. System calls provide an essential interface between a process and the operating system.

F. System Calls

Text-to-Speech (TTS) refers to the ability of computers to read text aloud. A TTS Engine converts written text to a phonemic representation, then converts the phonemic representation to waveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third-party publishers.

V. RESULTS

A. Speech to Text Module:

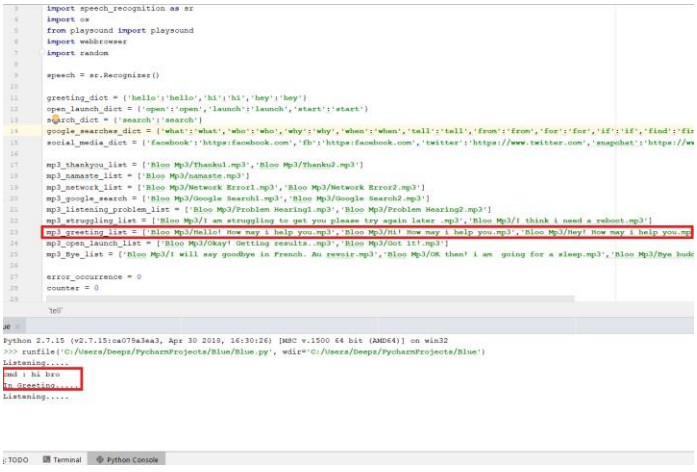


Fig.4. Implementation of Speech to text Module

As we have given the “hi bro” input in the form of speech and the speech recognition module converted it into the text format and the output give back to the user was “Hello/Hey/Hi! How may i help you”. AS shown in fig.4.

B. Text to Speech Module:

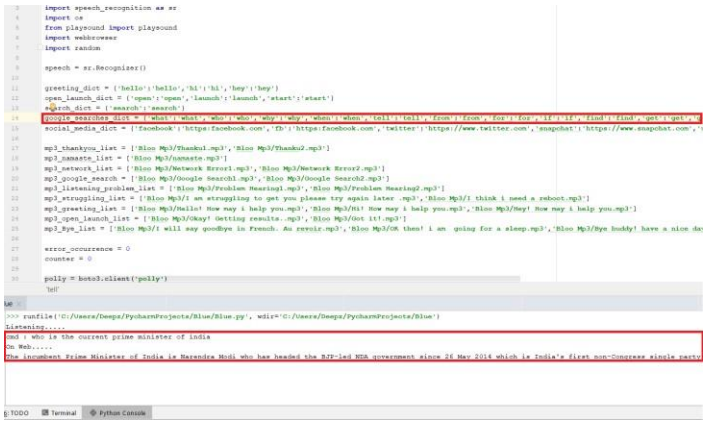


Fig. 5. Implementation of Text to Speech Module

In the text to speech test the input was give “who is the current prime minister of india” in the form of speech input then the speech format in converted in text format and the result is searched on the web then the result is give in the format of speech to the user as shown in fig.5.

C. System Call Module:

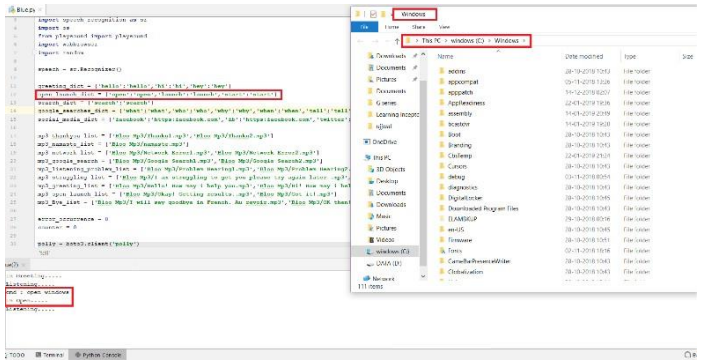
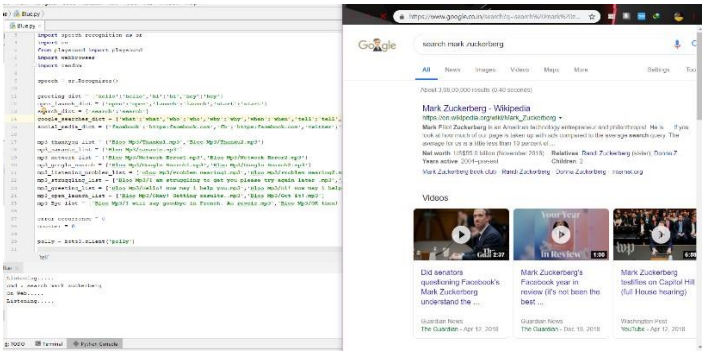


Fig. 6. Implementation of System Call Module

In this module, we successfully integrated the TTS and STT system. The output was in the form of performing the System task as open/start/launch in the hard disk. As shown in fig.6.

D. Google Search Module:

Using this feature we were successful in obtaining all the URL having any correlation with the input word. The input here was “Search Mark Zuckerberg” and the output was given in the form of web search in the following web browser by automatically getting



the link for the query given by the user to the assistant. Thus all links that would be obtained on a regular google search was given to the user. As shown in fig.7.

Fig.7. Implementation of Google Search Module

VI. CONCLUSION

In this paper, we discussed the design and implementation of a Digital Assistance. The project is built using open source software modules with PyCharm community backing which can accommodate any updates in the near future. The modular nature of this project makes it more flexible and easy to add additional features without disturbing current system functionalities.

It not only works on human commands but also give responses to the user on the basis of query being asked or the words spoken by the user such as opening tasks and operations. It is greeting the user the way user feels more comfortable and feels free to interact with the voice assistant. The application should also eliminate any kind of unnecessary manual work required in the user life of performing each and every task. The entire system works on the verbal input rather than the text one.

VII. FUTURE SCOPE

The possibility of added functionality required in making the assistant more accurate and fast while the interaction with the user. This project can be further improved by implementing the voice command in Google search queries. Better speech recognition so that the user can get prompt output and applications such as locking pc or opening pc on the commands of the user. Form Filling Functionality: Sometimes user are facing trouble while filling the form by their own each and every time so there might be chances of adding the feature like saving user's data and when in need the forms get automatically filled by simple commands.

In coming days our proposed system can be applied in multilingual application so that a person can use the application in their own language without any trouble. In addition, our proposed system can be deployed with the IoT. In future our proposed system will be able interpret the textual description in a much better way. The Image recognition can be used with much more details about the image captured through the camera. Enhancement to this system can be done by adding the features of currency recognition.

REFERENCES

[1] G. Bohouta, V. Z. Këpuska, "Comparing Speech Recognition Systems (Microsoft API Google API And CMU Sphinx)", Int. Journal of Engineering Research and Application 2017, 2017.

[2] B. Marr, The Amazing Ways Google Uses Deep Learning AI.

[3] Artificial intelligence (AI), sometimes called machine intelligence https://en.wikipedia.org/wiki/Artificial_intelligence

[4] Cortana Intelligence, Google Assistant, Apple Siri.

[5] Hill, J., Ford, W.R. and Farreras, I.G., 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. Computers in Human Behavior, 49, pp.245-250.

[6] K. Noda, H. Arie, Y. Suga, T. Ogata, Multimodal integration learning of robot behavior using deep neural networks, Elsevier: Robotics and Autonomous Systems, 2014.

[7] "CMUSphnix Basic concepts of speech - Speech Recognition process". http://cmusphinx.sourceforge.netlwiki/tutorialconcepts

[8] Huang, J., Zhou, M. and Yang, D., 2007, January. Extracting Chatbot Knowledge from Online Discussion Forums. In IJCAI(Vol. 7, pp. 423-428).

[9] Thakur, N., Hiwrale, A., Selote, S., Shinde, A. and Mahakalkar, N., Artificially Intelligent Chatbot.

[10] Mohasi, L. and Mashao, D., 2006. Text-to-Speech Technology in Human-Computer Interaction. In 5th Conference on Human Computer Interaction in Southern Africa, South Africa (CHISA 2006, ACM SIGHI) (pp. 79-84).

[11] Fryer, L.K. and Carpenter, R., 2006. Bots as language learning tools. Language Learning & Technology.