JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Digital Voice Assistant-Vision[implementation]

Prof. S.V.Shinde
*Computer Departement*
*Pune District Education Association*
*COE Manjari Bk*
*BCU id :- 52201482373*
*Mail id:-sashsir@gmail.com*
Pune, India

M.M.Joshi
*Computer Departement*
*Pune District Education Association*
*COE Manjari Bk*
Pune, India
*ABC id:-206235116095*
maithilijoshi1605@gmail.com

B.S.Kadam
*Computer Departement*
*Pune District Education Association*
*COE Manjari Bk*
Pune, India
*ABC id:-520968815930*
bhaktikadam151@gmail.com

K.D.Darekar
*Computer Department*
*Pune District Education Association*
*COE Manjari*
Pune,India
*ABC id:-922708398968*
darekarkalyani999@gmail.com

N.A.Tamboli.
*Computer Department*
*Pune District Education Association*
*COE, Manjari*
Pune,India
*ABC id:-958112361268*
tambolinikita12@gmail.com

***Abstract-*** **This paper shows the working of a device based on implementation of a voice command system as an intelligent personal assistant. The services provided by the device depends on the input given in the form of voice command by the user and ability to access information from a variety of online sources such as weather, telling time or accessing online applications to listen to music. This Voice driven device uses Raspberry Pi as its main hardware. Speech to text engine is used to convert the voice command to simple text. Query processing is then applied using natural language processing (NLP) onto this text to interpret the intended meaning of the command given by the user. After interpreting the intended meaning, text to speech conversion is used to give appropriate output in the form of speech. This device might provide a platform to visually impair to do their day to day tasks more easily like listening to music, checking weather conditions. ASR (Automatic speech recognition) is a method recognizer for speech it forward the recognition speculate to the NLU.**

***Keywords: Virtual Personal Assistant, Natural Language Processing, Query Processing, Raspberry Pi, NLU.***

## I. Introduction

Digital assistant is computer program designed to assist a user by answering questions and performing basic tasks. To interact with a digital assistant, must use a wake word, which device uses to activate the digital assistant. Digital assistant uses advanced Artificial Intelligence, natural language processing and understanding and machine learning. AI to learn as they go and provide a prenasalised, conservational communication. Combining historical information such as purchase preferences, home ownership, location, family size, so on, algorithms can create data models that identify patterns of behaviour and then refine those patterns as data is added. Existing examples of digital assistant are Apple's Siri, Google assistant, Alexa etc. Digital assistant gathers real time insights, which business can use to continually improve the user's experience and learn about their customers and employees

## II. System Requirement

1. *Hardware Requirement :*

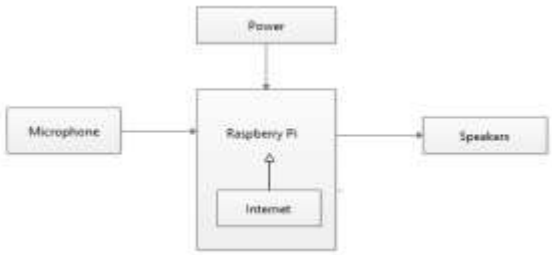
a. Microphone: The vocal commands given by the user which is used as input is given in through the microphone that is connected to the device. This vocal command is then later converted to simple text and keywords are searched through this text which helps the device to perform its functions and give out the expected results.

b. Raspberry Pi: Raspberry Pi is the major component of the device. It acts as a mini computer. It is indulged in all the activities since the beginning when the user gives the input till the end when the output is presented to the user. It sorts of binds all the components together. All the processing of the data takes place here.

c. Ethernet: The Ethernet cable helps us to provide the internet connection to the device. Internet plays a very important role in the operation of the device as it helps the device to do speech to text conversion, query processing through NLP and text to speech conversion. All these processes take place online that's why the internet connection is very essential.

d. Speaker: Speaker performs the last function in this process. The speaker helps the device to give out the output in the form of speech that is being converted from the text online. The speaker can be connected to the device through an AUX cable.

e. Putty: as a File Transfer Protocol PuTTY can connect to a remote machine through SSH. SSH (Secure shell) is a protocol that allows a secure connection. In this way, a PC can securely send and receive data from a remote server.



2. *Software Requirements:*

a. Input through microphone: The user gives in his command verbally to the device through the microphone in the form of input that the device later will process on. But before using the microphone it needs to be configured properly.

b. Speech to text conversion: The input that is given by the in the form of a vocal command is first converted to the plain text. This action is performed by using the Google speech recognition API. This API can recognize around 120 languages and variants so that it can support global user base. It’s easy to use and very effective.

c. Query Processing: This part of the whole process is the most important one. This process uses natural language processing (NLP) to operate. The input that is converted to the text by the device is studied in this step. Whole text is analysed, thereafter the tokens are identified from the text being received. By considering the tokens selected by the device it tries to interpret that what could be the action that the user wants it to perform. Once it analyses the intended action that user wants it to do it then performs the action that the user is expecting it to perform. Natural Language Toolkit (NLTK) is being used here to implement the natural language processing.

d. Text to speech conversion: Once the device interprets the intended meaning of the command given by the user it then performs the action and gives its output in form of text. For the user this text is then converted to the speech. For this conversion of the text to speech python text to speech package (pyttsx) is used.

e. Output through speakers: Once all the above steps are performed then it’s time to give the output to the user in the form of speech. This action is achieved by using the speaker. Speaker can be connected to the device by using an AUX cable.

## III. Voice Assistant(Vision)

First, Voice assistants are intelligent software that responds to voice commands and can run on any device, including smartphones, speakers, desktop/laptop computers, tablets, wearables, gaming consoles, TV consoles, virtual reality (VR) headsets, cars, and IoT devices. Examples include Amazon’s Alexa, Apple’s Siri, Google Assistant, and Microsoft’s Cortana.

This system is designed to be used efficiently on desktops. Personal assistant software improves user productivity by managing routine tasks of the user and by providing information from online sources to the user. VISION is effortless to use. Call the wake word ‘VISION’ followed by the command. And within seconds, it gets executed. Voice searches have dominated over text search. Virtual assistants are turning out to be smarter than ever. Allow your intelligent assistant to make email work for you. Detect intent, pick out important information, automate processes, and deliver personalized responses. This project was started on the premise that there is sufficient amount of openly available data and information on the web that can be utilized to build a virtual assistant that has access to making intelligent decisions for routine user activities

### *Artificial Intelligence*

Artificial Intelligence (AI) has been propelled into the mainstream of learning. AI has many areas like computer science, cognitive and learning sciences, game design, psychology, sociology, philosophy, mathematics, neuroscience, linguistics, defence industry, medicine and education. AI uses logical series of steps called algorithms and advanced cognitive computing technologies to use the techniques of search and pattern matching for providing solutions for the demanded answers. AI is an interdisciplinary field that is used for diagnosis of illnesses, criminal identification and artificial instructions. To develop communication between human and computer, AI possesses the ability to reason while processing a natural language and has different scope of data in terms of the developments in the above mentioned fields. AI, has some other descriptions as well. AI has the ability to comprehend, learn, solve, interpret and execute complex mental process. AI is a subfield of computer science. Natural Language Processing (NLP) is provided for human computer interaction in order to combine human learning and machine reasoning. NLP is the analysis of linguistic data, most commonly in the form of textual data such as documents or publications, using computational methods.

### *A. Technology used in Vision*

Vision uses Machine Learning technologies to function. Using ASR (Automatic speech recognition) to transcribe human speech (in this case, short utterances of commands, questions, or dictations) into text. Users speak natural language as voice commands in order to operate the mobile devices (all devices including phones, tabs) and its applications. The idea is to provide high level modelling primitives as integral part of a data model in order to facilitate the representation of real world situations and provide camera for more use of vision.camera in vision is basically used to detect motions and also recognize faces. It is also capable for authorization application.

### *B. System Architecture*

The architecture of digital voice assistant (Vision) designed with 3 layers as follows:

1. Client layer

Clients enable the user to access the digital assistant via voice with the following characteristics.

- Commonly Clients uses a microphone to capture the spoken input and a speaker is used to provide responses.

- As an extension to this Clients may also capture input from a specific modality recognizer.
- As an extension, Clients may also capture contextual information, e.g., location.
- As an extension a client may also receive commands to be executed locally.
- As an extension a client may also receive multimodal output to be rendered by a respective modality synthesizer.

2. Command Layer

This is service layer for a client to communicate with command management.
This layer has following characteristics:

- The IPA service acts as an interface between the IPA client and the command management and provider selection service.
- The output from the ASR is forwarded to the Provider Selection Service to get an appropriate meaning.
- Alternatively, the Service may receive multimodal or text input from the client and forwards it directly to the Provider Selection Service to determine meaning.
- Dialog Management receives recorded voice input from the Service and forwards it to the ASR
- Dialog Management makes use of the TTS to generate audio data to be rendered on the Client
- As an extension, it may also provide commands as output to be executed by the IPA Client
- As an extension Dialogs may also return multimodal output or text to be rendered by a respective modality synthesizer on the Client.
- The Dialog Manager is also responsible for a good user experience across the available Dialogs.
- The Command Manager follows the principle to fill in all slots that are known before prompting the user for it.
- The Command Manager also manages the session with a user. Conceptually, multiple sessions can be active in parallel. Commands are governed by Sessions, e.g., to free resources of ASR and NLU engines when a session expires.

- Linguistic phenomena, like anaphoric references and ellipsis are expected to work within a Session. The selected Digital Voice Assistant Provider or the Command Manager may have leading roles for this task.

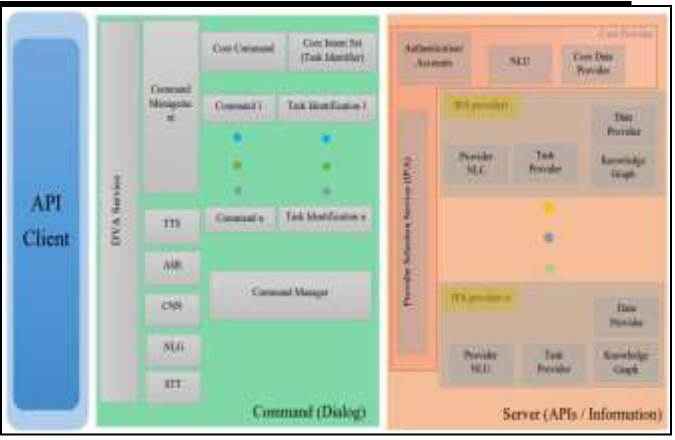


**Fig.1.Architecture of Vision**

- The Automated Speech Recognizer (ASR) receives audio streams of recorded utterances and generates a recognition hypothesis as text strings. Conceptually, ASR is a modality recognizer for speech.
- The Text-to-Speech (TTS) component receives text strings, which it converts into audio data. Conceptually, the TTS is a modality specific renderer for speech.

3. APIs/Data Layer

A service that provides access to all known Providers. This service also maps the task identifier Sets to the task identifier Sets in the command layer. It has the following characteristics:

- The Provider Selection Service receives input as text strings and returns results as task with all recognized entities from all Providers that are able to reply to the user input along with associated entities.
- In case the Provider Selection Service is called with preselected Providers only this one will be used.
- Providers the Accounts/Authentication to access them and optionally ASR and TTS capabilities can be added or removed as needed.
- The Provider Selection Service is stateless and always returns the responses.
- Providers along with an identification of the issuing Provider.
- The Provider Selection Service makes use of the Accounts/Authentication to access Provider.
- The Provider Selection Services maps the Provider task identifier Sets to the task Sets known by the Dialog Registry. The mapping must be configured when Providers are added.
- An NLU component that is able to extract meaning as tasks and associated entities from an utterance as text strings for Provider X.
- The Provider NLU may make use of the Data Provider to access local or internal data or access external services.
- The Provider NLU may make use of the Knowledge Graph to derive meaning.
- A knowledge graph to reason about the detected input from the Provider NLU and Data Provider to come up with some more meaningful results.

For better navigation of the user, it can open a map and thus helps in better accessibility.

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

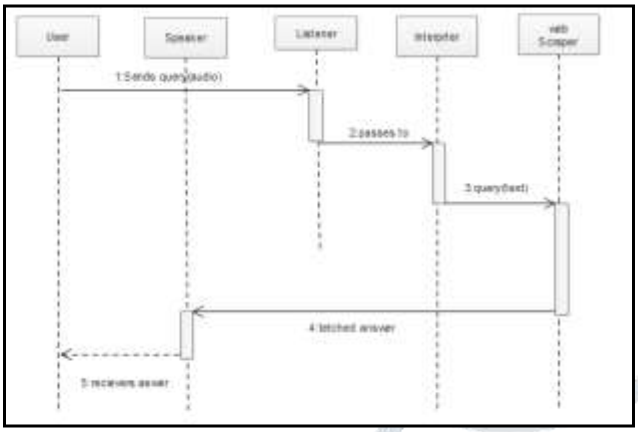


**Fig.2.Sequence Diagram**

*C. Working Of Vision*

Working of vision can be explained in following steps

- Voice Recognition – Whenever a person commands through his/her natural voice, the assistant must be able to convert that analog signal to digital one and then understand what was being said after concatenating the keywords altogether, and finally fixing/obeying the issue/command. This might sound trivial and easy but it is the first step towards reaching the next, since without overcoming the hurdle of country-wise accents, surrounding noises, and specific voices, one cannot successfully establish its working. It also timely learns how its user sounds while speaking specific words. The speech recognition that Vision uses is 80% accurate and has really low error rate.

- Send everything to the servers on the cloud – Vision does not work locally on a mobile device and eats its limited resources, but rather loads everything to the powerful computer servers so as to extend the maximum efficiency and continuously improvise. There is an algorithm that identifies the keywords and go down towards the flowchart branches (conceptually Tree data structure) that best match those keywords, so as to reach out to meaningful conclusions. Best case approach is used to solve such queries. If it fails, it searches for another branch. If it fails here too, it asks whether the user wants results from the Web. It hasn't reached to the point of conversational App but has numerous conditional statements in its coding that respond according to the user's action.

- Action based on what was commanded – Here is the most challenging thing. Vision or any other AI assistant you plan to develop must understand what you say. If it fails it might also drag you to potential dangerous situation. For instance, if you said to book a train, it must be capable to understand this and as well interact with other Apps to perform the given task. Plus it must not interact with those sites that aren't your interest, especially those that involve credit/debit card payment. One might get doomed if the assistant doesn't serve appropriately.

- The first step, speech to text, essentially converts voice command to a text input that your computer or smartphone gets from typing. Good 'speech to text' software like Apple Dictation, Google Docs voice typing and Dragon naturally speaking adjust for ambient noise and variation in voice tone/pitch/accent to provide accurate translation in multiple languages. Science Line explains how the software works:

- The software breaks your speech down into tiny, recognizable parts called phonemes — there are only 44 of them in the English language. It's the order, combination and context of these phonemes that allows the sophisticated audio analysis software to figure out what exactly you're saying … For words that are pronounced the same way, such as eight and ate, the software analyses the context and syntax of the sentence to figure out the best text match for the word you spoke. In its database, the software then matches the analysed words with the text that best matches the words you spoke."

- The second step, text to intent, interprets what exactly does the user mean. For example, if you say "tell me about Paris" in a conversational context, what should the Digital VI interpret as your real intent? Are you asking for latest news about Paris, or flight options to Paris, or current weather in Paris, or news stories about Paris Hilton? Web search engines solve this challenge by ranking answers to the 'query' in decreasing order of inferred intent. For Digital VI, the bar is higher as it has to abstract intent from a conversational input, and then respond with one best answer.

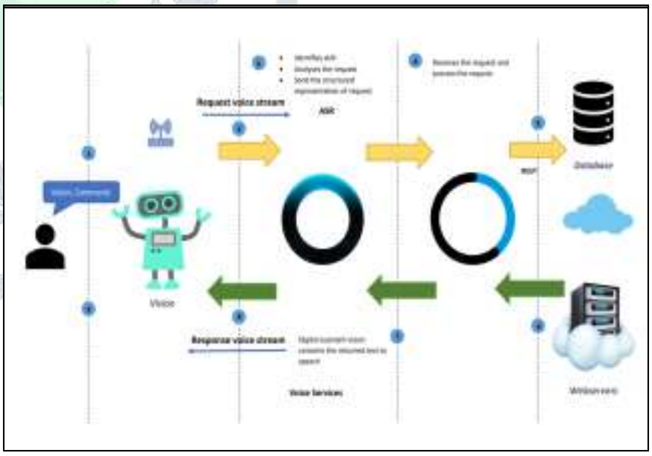


**Fig.3.Working of Vision**

## IV. Modules and Algorithm

1. Modules

In this proposed system, we have used the multi-model dialogue systems which process two or more combined user input modes, such as speech, head, and body movement in order to design the dedicated digital voice assistant system. We have modified and added some components in the original structure of general dialogue systems, such as ASR Model, Gesture Model, Interaction Model, User Model, Input-Output Model, Inference Engine, Cloud Servers, and Knowledge Base. The following is the structure of the Digital Voice Assistants:

A. Knowledge Base

There are two knowledge bases. The first is the online and the second is local knowledge base which include all data and facts based on each model, such as facial and body data sets for gesture modal, speech recognition knowledge bases, dictionary and spoken dialog knowledge base for ASR modal, video and image body data sets for Graph Model, and some user's information and the setting system

.

B. Movement / Gesture Model

The Gesture Model analyzes video and image in real-time by using the Gesture Model and extracts frames of the video that collect by the camera and the input model; then it sends those frames and images to the Gesture Model and applications in Cloud Servers for analyzing those frames and images. This model also uses the camera in the input model to read the movements of the human body then it sends all data to the gesture model and applications in Cloud Servers to analyze those frames and images and returning the result.

C. ASR (Speech Recognition) Model

The speech recognition model will work in real-time with the microphone in the input model with the ASR model in Cloud Servers to recognize the utterances that a user speaks into a microphone and then convert it to text; then it sends the text to the applications in Cloud Servers to analyze the text and returning the result.

D. Interaction Model

This is the main model that will be used to provide interaction between users of the system and the system models by receiving the data from the input model and analyzing the data to send for each model based on its tasks, then returning result that will be used to make the final decision.

E. Inference Engine

The inference engine works together with the Interaction Model in the chain of conditions and derivations and finally deduces the outcome. they analyze all the facts and rules, then sorts them before concluding to a solution.

F. User Model

This model has all information about the users that will use the system. It can include personal information such as users' names and ages, their interests, their skills and knowledge, their goals and plans, their preferences and their dislikes or data about their behavior and their interactions with the system. All information will be collected by asking the user some questions then storing all answers in the Knowledge Base.

G. Input-Output Model

This model will organize the work of all input devices that the system uses to collect the different data from microphone, camera and Kinect. Also, this model includes intelligence algorithms to organize the input information before sending the data to the Interaction Model. This model will receive the final decision from the Interaction Model with an explanation, then it will choose the perfect output device to show the result such data show, speakers or screen based on the result.

2. Algorithm

A speech recognition algorithm or voice recognition algorithm is used in speech recognition technology to convert voice to text. To convert text to speech, the ML system must perform the following:

Convert text to words:

Firstly, the ML algorithm must convert text into a readable format. The challenge here is that the text contains not only words but numbers, abbreviations, dates, etc. These must be translated and written in words. The algorithm then divides the text into distinct phrases, which the system then reads with the appropriate intonation. While doing that, the program follows the punctuation and stable structures in the text. Each sentence can be pronounced differently depending on the meaning and emotional tone. To understand the right pronunciation, the system uses built-in dictionaries. If the required word is missing, the algorithm creates the transcription using general academic rules. The algorithm also checks on the recordings of the speakers and determines which parts of the words they accentuate.

The system then calculates how many 25 millisecond fragments are in the compiled transcription. This is known as phoneme processing.

A phoneme is the minimum unit of a language's sound structure.

The system describes each piece with different parameters: which phoneme it is a part of, the place it occupies in it, which syllable this phoneme belongs to, and so on. After that, the system recreates the appropriate intonation using data from the phrases and sentences. Finally, the system uses an acoustic model to read the processed text. The ML algorithm establishes the connection between phonemes and sounds, giving them accurate intonations. The system uses a sound wave generator to create a vocal sound. The frequency characteristics of phrases obtained from the acoustic model are eventually loaded into the sound wave generator.

## V. IMPLEMENTATION

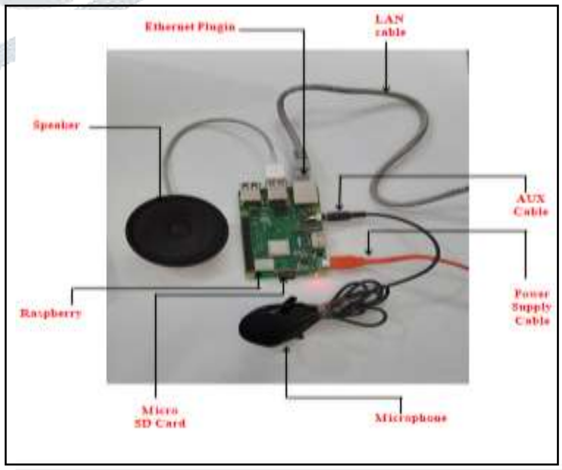


**Fig.4. Working Model**

The device that is used as virtual personal assistant uses Raspberry Pi as its major component. Raspberry Pi can be thought of as a small and affordable computer. It has ports on it through which other components of the device like speaker and mic can be connected to it. These components are then used to take in command from the user as input and then give out the results as outputs.

The device takes in the command from the user through the microphone being plugged into the raspberry pi in the form of input. The user gives its command in his natural language to the device. The vocal command is converted to plain text by speech to text synthesis .The device then performs query processing using natural language processing (NLP) on the plain text derived from the command given by the user. During query processing the device tries to interpret the intended meaning of the instruction given by the user. The keywords are searched in the sentence and the data that seems irrelevant to the device is ignored. Device figures out intentions of the user by using these keywords from the sentence. After the command is interpreted device simply performs the given task that user asked it to do. The performance of the device varies in different situations. Sometimes the input that the user gives through his vocal commands is not taken in or considered by the device. One reason for this is that the distance between the mouth of the user and the microphone is very large. Another reason maybe that the environment in which the user is giving the command to the device is very noisy. Also sometimes the device may give out some output that the user was not expecting, this problem may arise due to the unusual accent used by the user.

## VI. Conclusion and Future Scope

In the near future, voice assistants are also expected to take a more proactive role. Rather than just waiting for user commands, assistants will collect context-specific information and then take the initiative by making helpful suggestions to the user. For example, people can interact with their in-car voice assistants to get information about fuel levels, diagnostics, and service needs or system settings that may need adjustment. So when fuel levels are low, the voice assistant may suggest going to the nearest gas station (with GPS directions if needed).

The future of voice search and assistants is looking bright. With the number of people already seeing how convenient those tools can be and the growing number of devices that use voice recognition. It's clear that the technology will soon be everywhere, and with 5G and improvements in machine learning, voice assistants might at some point become tools we can’t live without.

## Acknowledgment

We would like to thank our guide Mr. S.V. Shinde for providing us with useful and valuable comments on the study. This study would not have been possible without their support. All remaining errors, if any, are our responsibility. The usual disclaimer applies

## Refrences

[1] R. Belvin, R. Burns, and C. Hein, “Development of the HRL route navigation dialogue system,” in Proceedings of ACL-HLT, 2001, pp. 1–5.

[2] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T.J.Hazen,and L.Hetherington, “JUPITER: A Telephone Based Conversational Interface for Weather Information,” IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, pp. 85–96, 2000.

[3] M. Kolss, D. Bernreuther, M. Paulik, S. St¨ucker, S. Vogel, and A. Waibel, “Open Domain Speech Recognition & Translation: Lectures and Speeches,” in Proceedings of ICASSP, 2006.

[4] D. R. S. Caon, T. Simonnet, P. Sendorek, J. Boudy, and G. Chollet, “vAssist: The Virtual Interactive Assistant for Daily Homer-Care,” in Proceedings of pHealth, 2011.

[5] Crevier, D. (1993). AI: The Tumultuous Search for Artificial Intelligence. New York, NY: Basic Books, ISBN 0-465-02997-3.

[6] Sadun, E., &Sande, S. (2014). Talking to Siri: Mastering the Language of Apple's Intelligent Assistant. Que Publishing. [6] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley. “Music, Search, and IoT: How people (really) use voice assistants,” ACM Trans. Comput. - Hum. Interact., vol. 26.3, pp. 1-28, 2019.

[7] Blog.google, ‘This Year’s Founders’ Letter’, 2016. [Online].Available: https://www.blog.google/topics/inside-google/thisyearsfounders-letter/ [Accessed: 15- Oct-2020].

[8] S. Zuboff, “Big other: surveillance capitalism and the prospects of an information civilization,” J. Inf. Technol., vol. 30.1, pp. 75-89, 2015.

[9] Saadman Shahid Chowdury, Atiar Talukdar, Ashik Mahmud, Tanzilur Rahman Domain specific Intelligent personal assistant with bilingual voice command processing IEEE 2018.

[10] Polyakov EV, Mazhanov MS, AY Voskov, LS Kachalova MV, Polyakov SV $^3$Investigation and development of the intelligent voice assistant for the IOT using machine learning Moscow workshop on electronic technologies, 2018.

[11] Khawir Mahmood, Tausfer Rana, Abdur Rehman Raza Singular adaptive multi role intelligent personal assistant (SAM-IPA) for human computer interaction International conference on open-source system and technologies,2018.

[12] M. McTear .2016. The Dawn of the Conversational Interface. Springer International Publishing Switzerland 2016 [8] CM Research: High quality research requires investment 2016.

[13] Y. Nung, A. Celikyilmaz.Deep Learning for Dialogue Systems. Deep Dialogue...

[14] Amazon. Amazon Lex is a service for building conversational interfaces. https://aws.amazon.com.

[15] Microsoft. Cortana Intelligence. https://azure.microsoft.com.

[16] B. Marr. The Amazing Ways Google Uses Deep Learning AI. https://www.forbes.com..

[17] K. Wagner. Facebook's Virtual Assistant 'M' Is Super Smart. It is Also Probably a Human. https://www.recode.com.