**Personal Voice Assistant**

**By:**

**Pulkit Gupta (16BCE0373)**

**Human Computer Interaction**

(CSE4015)

**Slot: C2+TC2**

**Professor: Ananda Kumar S**



Vellore-632014, Tamil Nadu, India

# **Abstract**

Voice interaction has become a feature in many commercial devices such as mobile phones and tablets. More recently, voice has become the primary interface with standalone screen less devices such as the Amazon Echo and Google Home. These interfaces, often referred to as voice user interfaces (or VUIs), conversational agents, or intelligent or virtual personal assistants, are described as embodying the idea of a virtual butler that helps you get things done.

Researcher’s adoption of such technologies as “conversational interfaces” resonates in many ways with the advertised user experience of such devices: specifically, as technologies that it is possible to ‘have a conversation’ with and ‘just ask’ questions of. In addition, some VUIs are pitched as being especially suited to use in the home for a variety of purposes:

-to help with cooking

-play music

-Access news and information

-or play games with.

In this project we will take user voice as the input and then process the natural human language into the machine-readable language and then the query is solved by the system and the output is given back to the user in natural language on random selection basis from the user.

**Keywords:**

Artificial intelligence, Hidden Markov Model, Dialogs, Training phase, Intents

**Introduction**

A **virtual assistant** is a software agent that can perform tasks or services for an individual. Sometimes the term "chatbot" is used to refer to virtual assistants generally or specifically those accessed by online chat (or in some cases online chat programs that are for entertainment and not useful purposes).

As of 2017, the capabilities and usage of virtual assistants are expanding rapidly, with new products entering the market. An online poll in May 2017 found the most widely used in the US were Apple's Siri (34%), Google Assistant (19%), [Amazon Alexa](https://en.wikipedia.org/wiki/Amazon_Alexa) (6%), and Microsoft Cortana (4%).[[1]](https://en.wikipedia.org/wiki/Virtual_assistant#cite_note-1) Apple and Google have large installed bases of users on [smartphones.](https://en.wikipedia.org/wiki/Smartphone) Microsoft has a large installed base of Windows-based personal computers, smartphones and smart speakers. Alexa has a large install base for smart speakers.

The first tool enabled to perform digital speech recognition was the IBM Shoebox, presented to the general public during the 1962 Seattle World's Fair after its initial market launch in 1961. This early computer, developed almost 20 years before the introduction of the first IBM Personal Computer in 1981, was able to recognize 16 spoken words and the digits 0 to 9. The next milestone in the development of voice recognition technology was achieved in the 1970s at the Carnegie Mellon University in Pittsburgh, Pennsylvania with substantial support of the United States Department of Defense and its DARPA agency. Their tool "Harpy" mastered about 1000 words, the vocabulary of a three-year-old. About ten years later the same group of scientists developed a system that could analyse not only individual words but entire word sequences enabled by a Hidden Markov Model[.[3]](https://en.wikipedia.org/wiki/Virtual_assistant#cite_note-spracherkennung1-3) Thus, the earliest virtual assistants, which applied speech recognition software were automated attendant and medical digital dictation software.

[[4]](https://en.wikipedia.org/wiki/Virtual_assistant#cite_note-4) In the 1990s digital speech recognition technology became a feature of the personal computer with Microsoft, IBM, Philips and Lernout & Hauspie fighting for customers. Much later the market launch of the first smartphone IBM Simon in 1994 laid the foundation for smart virtual assistants as we know them today.[[5]](https://en.wikipedia.org/wiki/Virtual_assistant#cite_note-5) The first modern digital virtual assistant installed on a smartphone was Siri, which was introduced as a feature of the iPhone 4S on October 4, 2011.[[6]](https://en.wikipedia.org/wiki/Virtual_assistant#cite_note-6) Apple Inc. developed Siri following the 2010 acquisition of Siri Inc., a spin-off of SRI International, which is a research institute financed by DARPA and the United States Department of Defence.

The **objective** of this project is to provide help to the user with almost all the basic tasks an actual assistant can do in a short period of time with conservation of man power.

These tasks include playing music, adding notes, fixing appointments, checking calendar, accessing location and giving you the weather details, providing you with the latest news, giving latest sports scores, calling, giving directions using maps, recommending movies, adding reminders, sending emails, providing stock information, setting up alarms, finding contacts, adding notes, doing web searches, calling an über, reading WhatsApp messages and sending WhatsApp messages and opening apps.

Using these functionalities, the life of a person is surely to get easier and smoother. It’s built on the platform of google assistant so it can easily connect to your Bluetooth enabled car and you can easily answer and reject calls while driving.

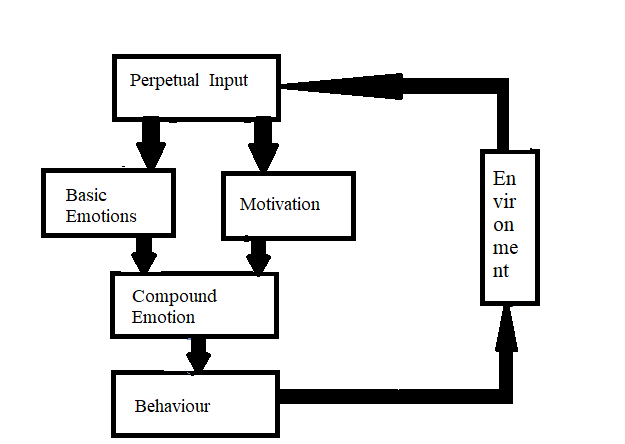
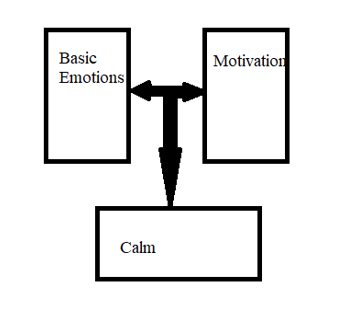


Figure: System structure chart

**Literature Review/ Related Work**

[Paper1] This paper was a very well written paper with certain that were very significantly highlighted which included the issues in the making of the voice driven interfaces that is the virtual assistant in the complex applications such as using the computers to do a complete scan of the computer checking each and every file etc. However, as there were problems, the paper produced the solution with it too. The solutions included the approach that have tested the observation of the user interaction by maintaining a spreadsheet which recorded all the communications taking place in the environment through the use of the voice command interface.

[Paper2] This paper written by Nicholas Mulhern and his colleagues is inclined towards solving the issue that the handicapped people face in the normal day to day life. Some people have neuromuscular disease due to which they can’t carry out common day to day tasks. In order to solve this issue, the authors thought of making a device application which is a modified version of offline control of the household device using the input and the output interface processor which is built as an application in the smartphones like Samsung galaxy S. This application has made this disabled people more independent from the other people who were required to take care of these handicapped in the ancient times. However, limitations are there such as the compatibility issue, as this application is supported by few devices only. The paper ensures future developments.

[Paper3] This paper had a combination of various technologies together like the voice recognition, the facial expression. These features combined with the learning attribute of the virtual assistant made the assistant more powerful emotionally. The assistant would be able to connect with the user emotionally by looking at the facial expressions and the tone in which the user gave input. Slowly as the assistant faced different emotions conditions, it started to learn about them so that it could provide the optimized solutions that could be helpful. It had combination of 8 emotions prototype based upon which the psychological model of the user was produced. This technology however, cannot be always relied as there can be a case when there is a new situation that the device has to face, and it doesn’t know how to react to it and it may lead to a wrong decision. This technology is still at its developing stage.

[Paper4] This paper is completely based upon the home automation system. In this system, the home is installed with the appliances or the electrical devices which take in the voice command to either start or stop working. This technology can be useful for the handicapped people as discussed in paper [3]. This is a very cheap technology that can be implemented in variety of voices and manners of speaking. However, the system lacks the only thing that it was tested on a very small platform and lacks its performance results on the bigger platforms such as the villas. Further development is recommended in this paper.

[Paper5] This paper is the better part of the paper [4]. But it lacks some capabilities too. The working of this proposed model in this paper is given by the different sensors giving the status of the appliance that is plugged into the wall outlet, whether it’s on or off. The user then may control the devices using the voice commands comfortably. This again is useful for the people how are handicapped or demand comfort. However, the limitation is that the environment if noisy will interfere with the commands given and the output won’t be an accurate one. For this working, a very calm and silent environment is needed. This paper but overcomes the situation of a big space because the user when plugs in the device gets connected a centralized system (e.g. SCADA). So, if user wishes to operate in villa using voice commands, he/she can easily do that.

[Paper 6] By collecting and studying audio data from month-long deployments of the Amazon Echo in participants’ homes—informed by ethnomethodology and conversation analysis what they did is that they studied the documents, the methodical practices of VUI users, and how that use is accomplished in the complex social life of the home. Data that they present shows how the device is made accountable to and embedded into conversational settings like family dinners where various simultaneous activities are being achieved. They discuss how the VUI is finely coordinated with the sequential organisation of talk. Finally, they locate implications for the accountability of VUI interaction, request and response design, and raise conceptual challenges to the notion of designing ‘conversational’ interfaces.

[Paper 7] In this paper, they present an alternative approach: the development of a strategy that minimizes the number of errors visible on the user interface, given a fixed speech recognizer performance. They introduce a unified view on the recognition act, i.e. the classification of an incoming sound pattern, and derive a self-optimization algorithm for the user interface. With this algorithm, each recognition act is optimized with respect to its own criterion, independently from all other recognition acts. As a result, the voice control system adapts flexibly to the demands of the respective situation.

[Paper 8] Optimism is growing that the near future will witness rapid growth in human-computer interaction using voice. System prototypes have recently been built that demonstrate speaker-independent real-time speech recognition, and understanding of naturally spoken utterances with vocabularies of 1000 to 2000 words, and larger. Already, computer manufacturers are building speech recognition subsystems into their new product lines. However, before this technology can be broadly useful, a substantial knowledge base is needed about human spoken language and performance during computer based spoken interaction. This paper reviews application areas in which spoken interaction can play a significant role, assesses potential benefits of spoken interaction with machines, and compares voice with other modalities of human-computer interaction. It also discusses information that will be needed to build a firm empirical foundation for the design of future spoken and multimodal interfaces. Finally, it argues for a more systematic and scientific approach to investigating spoken input and performance with future language technology.

[Paper 9] VoiceXML is a standard language for developing voice based applications. VoiceXML applications have more advantages over traditional Interactive Voice Response (IVR) systems because they can be used through any type of phones and also accessed via a computer. Voice User Interface (VUI) design is an integral part of developing any VoiceXML application. In this paper, the VUI for a VoiceXML ‘Cinema Service’ telephone application is designed and a number of experiments are undertaken to help the design of the VUI. The experiments focus on users’ navigation, memory and age group, and preferences. Conclusions are drawn based on the experiments for future design and development.

[Paper 10] this paper describes the experiences designing a voice interface in rural India. It outlines the design process from initial contextual inquiry to a formal user evaluation, and use this discussion to motivate research guidelines for others designing voice interfaces in developing regions. The three guidelines are to build around existing information systems, to iterate on the design through user testing, and to explore design alternatives through empirical analysis. It also share some practical lessons learned in designing, implementing, and evaluating information systems for developing regions in general.

**Survey table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.no | Author | Parameters Satisfied | Advantages | Disadvantages | Year of publication |
| 1 | [11]Alexander I. Rudnicky | The issues that arise in building voice driven interfaces to complex applications are there in this paper and describes some of the approaches that they have developed for this purpose. | To test the approaches, the have implemented a voice spreadsheet and have begun observation of users interacting with it. | There was no stable picture of the relationship between the different conditions | 2010 |
| 2 | [12]Nicholas Mulhern, Neil McCaffrey, Nicholas Beretta, Eugene Chabot PhD, Ying Sun | Utilizing assistive technology to aid persons with neuromuscular disabilities, this study improves upon previously created offline voice control of household devices through an input-output peripheral interface control processor driven using an application on the Galaxy S Android smart phone. | People with physical disability now have an increased ability to operate within a household environment independently. | Works only on some of the smartphones | 2012 |
| 3 | [13]Zhiliang Wang, Ning Cheng, Yumei Fan, Jiwei Liu, and Changsheng Zhu | The purpose of this paper is to construct a virtual assistant. The compound emotion consists of eight prototype basic emotions. Based on this theory, a psychology model is constructed. The parameters are adjusted to simulate different human psychologies. A real time facial expression and voice recognition and synthesizing technology are combined to construct a virtual assistant. | It is more interactive since it emotionally connects with the user | It can’t be relied always as it is emotion and facial expression driven. | 2005 |
| 4 | [14]Rahul Biswash, Sandeep Nagar | Home automation refers to automatic manipulation of home appliances. In present paper, we present a low cost open sourced scheme to implement home automation of basic electrical devices which can be switched on and off using audio commands. To test the scheme, a model setup of a home has been designed and experiments on the setup proves the reliability of scheme with successful implementation of a variety of voices and manners of  speaking | It is cheap | The system is only tested on small model | 2015 |
| 5 | [15]Aqueel-ur-Rehman, Royda Arif, Hira Khursheed | The developed system can be integrated as a single portable unit and allows one to wirelessly control lights, fans, air conditioners, television sets, security cameras, electronic doors, computer systems, audio/visual equipment’s etc. and turn ON or OFF any appliance that is plugged into a wall outlet, get the status of different sensors and take decision accordingly. | A centralized control unit for healthy and comfortable lifestyle to the users, and also for the sick or handicapped and people living alone. Everyone can easily handle all their tasks at a convenience. | System sometimes in noisy environment produces error and thus requires absolute silence. | 2014 |
| 6 | [17]Martin Porcheron, Joel E. Fischer, Stuart Reeves and Sarah Sharples | Coordination of the voice user interface with the sequential organization of talk is discussed in this paper. | Data from amazon echo is collected and studied. | Different amazon echo were used but they are costly | 2018 |
| 7 | [18]Wolfgang Tschirk | This paper presents an approach for the development of a strategy that minimizes the number of errors visible on the user interface, given a fixed speech recognizer performance | The voice control adapts flexibly to the demands of the respective situation. | However, it does not adapts in all situations. | 2004 |
| 8 | [19]Philip R. Cohen and Sharon L. Oviatt | This paper reviews application areas in which spoken interaction can play a significant role, accesses potential benefits of spoken interaction with machines | Reliable method to predict when voice input will be the most effective, efficient or preferred modality of communication | It is sometimes difficult to understand some of the naturally spoken languages | 2007 |
| 9 | [20]Daniel Mecanovic, Hao Shi | The experiments focus on users’ navigation, memory and age group, and preferences. Conclusions are drawn based on the experiments for future design and development | The voiceXML can be used through any type of phone and through a computer also. | Not able to provide reliable output if user personality inputs are not provided correctly or if some other user is using the phone. | 2005 |
| 10 | [21]Neil Patel, Sheetal Agarwal, Nitendra Rajput, Arun Kumar, Amit Nanavati, Paresh Dave, Tapan S Parikh | This project iterate on the design through user texting and to explore design alternative through empirical analysis. | The voice interface is actually helping farmers with limited knowledge and thus improving the status of our country. | Sometimes does not allow useful heuristics for practitioners | 2008 |

## **Innovation component in the project**

The most innovative part of the project is that it can work on any platform, be it any mobile OS like android, iOS, windows and blackberry. Most of the virtual assistants are OS specific, for example, Siri is only available for Apple devices, i.e., iOS, Google on Tap is for android devices, Microsoft Cortana is just for windows phones and Blackberry Assistant is just available for Blackberry devices. This project is compatible for every platform.

Violet is filled with humour to lighten the mood and help people relax. She cracks great jokes and tend to make people laugh which, at the end makes her likeable and more preferred compared to her competitors.

Unlike other voice assistants, Violet uses open source python libraries which have a lesser tendency to fail and return requests a lot faster and in a smoother fashion.

Violet also connects to car that only support Apple Car Play, and Android Car Support which makes her more versatile compared to the others, for example, Siri connects to the cars that only support Apple Car play, Bixby - The voice assistant made by Samsung is compatible with Android enabled Cars only, and so on. But Violet is available for all the cars.

### Tools used

Tools used to build Violet are as follows:

1. Xcode

2. Python

3. Terminal

4. Pytts

5. Mpg321

**Proposed Methodology**

To achieve good classification accuracy, it’s important to provide your agent with enough data. The greater the number of natural language examples in the **Training Phrases** section of **Intents**, the better the classification accuracy.

When you create a new intent, start with examples that have the number parameters. This way you will define what entities should be used in this intent and name all the parameters the right way.

When you save an intent, Dialog flow will begin training your agent with the new data you've added. Until the training is complete, the updates may not be reflected in the agent.

As you and your users chat with your agent, you can access the conversation logs by clicking Training in the left side menu.

The logs are presented in two views:

* Training - This view shows conversations with the agent for review and performance improvements. Each user request is a list item, showing the intent that will be used for processing, as well as the current parameter annotation. You can reassign inputs to correct intents and fix annotations. Every time you approve changes, the agent is trained, and the results in the tab are updated.
* History - This view displays the conversations in a plain mode. This way you can see latest conversations with your agent in chronological order.

Assigning an input to an intent adds the example as a **Training Phrases** entry for that intent. Training your agent using this method is good for adding specific examples from users' interactions. Disabling interaction logs Depending on your agent, conversation logs could include personally identifiable or confidential information and your agent may need to comply with legal or other restrictions.

Dialogs

There are two types of dialogs to consider when building voice interaction scenarios:

* **Linear dialogs** - the aim of which is to collect the information necessary to complete the required action (e.g. find the best hotel, turn on the right light bulb, or play the desired song)
* **Non-linear dialogs** - which may have several branches, depending on users’ answers

Linear Dialog

To build linear dialogs, we use a feature known as slot filling. The following example is an agent for a hotel booking app.

Since our app understands natural language, our travellers can request a hotel in many ways. Depending on how they word their request, there may be several different search parameters:  *• “I need to book a hotel."*

* *“Find a cheap hotel in Rome.”*
* *“Book a hotel in New York, check in November 10, check out November*

*15.”*

Let’s say our app requires at least three parameters to start the search:

* destination
* check-in date
* check-out date

This is where the slot filling feature comes in. Even before you start listing examples in the **Training Phrases** section, you can list all of the parameters your app can use to perform a search, both required and optional.

When you mark a parameter as **required**, you’ll be asked to write prompts that your app will address to your user when they make a request that does not include that parameter.

You can change the order in which your agent will ask these questions by dragging and dropping the parameters in the list.

Your agent will continue to ask these prompts until it has collected information for all required parameters. At any time, users can say "Cancel" and start from the beginning.

Non-linear Dialogs

Let’s look at an example of a non-linear dialog. In this example, our agent handles a customer satisfaction survey for a hotel.

It starts with these two questions:

* How would you rate the location of the property?
* How would you rate the facilities at the hotel?

For each question, there are four accepted answers:

* poor
* fair
* good
* excellent

To build this dialog we’ll need to use contexts. Here’s how it’s done:

First, we create an intent which reacts to the “start” command and triggers the dialog. In this intent we’ll ask the first question and set the outgoing context as “location-question”. As a result, the preferred intents for the answers to this question will have this as the incoming context.

Next we create intents for each of the four expected response variations, with the incoming context “location-question”. These four questions will work only while this context is active.

Now we need setup moving to the next question. To do this we make the following changes to the intents we just created:

* Remove "location-question" as the output context and add "facilitiesquestion"
* Set the **Action** to question. Location
* Add a parameter with the name as "Rate" and the value as the corresponding rating (poor, fair, good, excellent)
* Add a text response for the next question ("How would you rate the facilities at the hotel?”)

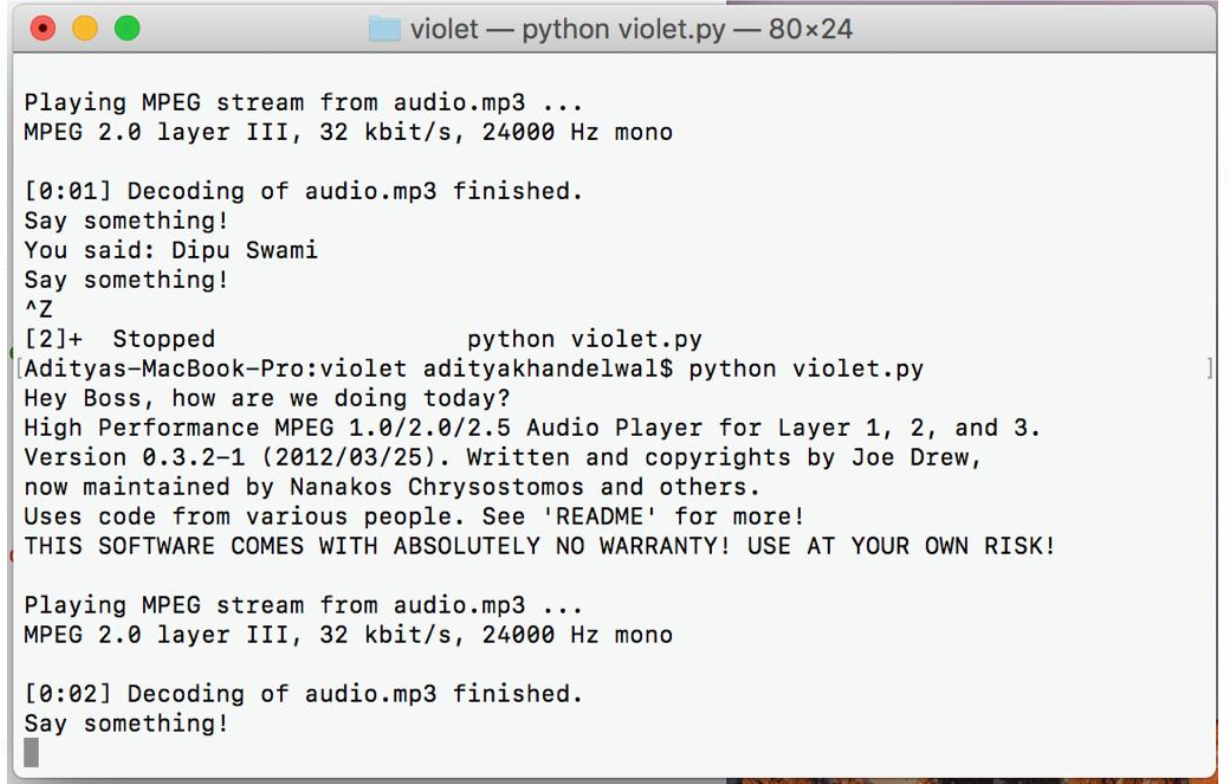
Once again, we mark the action and parameters we want to receive when these new intents are matched, and add the final response in the **Speech Response** field.

**Performance analysis using graphical approach**

We have analysed the performance (**queries answered correctly**) of Violet on different dates after we have developed the assistant and as the time passes by the assistant keeps on learning based on the training given to the training.

**Results**

The virtual assistant answer the queries of the user based on the question asked by the user and the data stored in the data list and the output is given in the form of audio.



### Conclusion

Violet is a personal assistant that can complete all most all the basic tasks like fixing your calendar, giving you the weather details, providing you with the latest news, giving latest sports scores, calling, giving directions using maps, recommending movies, adding reminders, sending emails, providing stock information, setting up alarms, finding contacts, adding notes, doing web searches, calling an über, reading WhatsApp messages and sending WhatsApp messages and opening apps.

It connects seamlessly with your smart wearable’s and cars Bluetooth and also your Bluetooth devices like earphones and headphones.

The software is working perfectly with each and every hardware required, all thanks to the libraries provided by python.

**Future work**

We will make the Violet to learn more with the time and answer more queries of the user correctly and try to reduce the bugs as far as possible.

We will also try to make Violet also available for smart wearables like smart watches and smart glasses and many other applications which will reduce human efforts and do most of the things by speaking with the Violet.

### References

1. <https://dialogflow.com/docs/dialogs>
2. <https://dialogflow.com/docs/agents>
3. <https://dialogflow.com/docs/intents>
4. <https://dialogflow.com/docs/entities>
5. <https://dialogflow.com/docs/actions>
6. <https://dialogflow.com/docs/rich-messages>
7. <https://dialogflow.com/docs/contexts>
8. <https://dialogflow.com/docs/analytics>
9. <http://journals.sagepub.com/doi/full/10.1177/2332858417749220>
10. [https://www.fastcompany.com/3059719/handicapping-the-ai-assistants-from-siri-toozlo](https://www.fastcompany.com/3059719/handicapping-the-ai-assistants-from-siri-to-ozlo)
11. <http://www.aclweb.org/anthology/H89-1015>
12. <http://www.ele.uri.edu/courses/bme484/2012/NEBC_Android.pdf>
13. <https://link.springer.com/chapter/10.1007/11573548_74>
14. <https://www.researchgate.net/publication/283495650_Low_Cost_Voice_Activated_Home_Automation>
15. <https://www.researchgate.net/publication/270271814>
16. <http://journals.sagepub.com/doi/full/10.1177/2332858417749220>
17. <http://www.cs.nott.ac.uk/~pszsr/files/porcheron-2018-voice-interfaces-in-everyday-life.pdf>
18. <https://ieeexplore.ieee.org/document/7079662/>
19. [http://www.pn as.org/content/pnas/92/22/9921.full.pdf](http://www.pnas.org/content/pnas/92/22/9921.full.pdf)
20. <https://link.springer.com/chapter/10.1007/978-3-540-31849-1_106>
21. <http://hci.stanford.edu/neilp/pubs/slt2008.pdf>