Implementation of Voice User Interface using Speech Recognition

Pranali Joshi
Department of Computer Engineering
Dr.D.Y.Patil College Of Engineering,
Ambi, Pune

Ravi Patki
Department of Computer Engineering
Dr.D.Y.Patil College of Engineering,
Ambi, Pune

ABSTRACT

A voice–user interface (VUI) makes human interaction with computers possible through a voice/speech platform in order to initiate an automated service or process. A VUI is the interface to any speech application. Controlling a machine by simply talking to it was science fiction only a short time ago. Until recently, this area was considered to be artificial intelligence. However, with advances in technology, VUIs have become more commonplace, and people are taking advantage of the value that these hands-free, eyes-free interfaces provide in many situations.

The system will be implemented by using techniques such as speech and language processing, human language technology, natural language processing, computational linguistics, and speech recognition and synthesis. The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.

These voice applications can provide intrinsically comfortable, easy-to-use, and efficient way for users to interact with computer. And as user can use commands in his mother tongue so it is asking your friend computer to do something for you. As a technology for expression, voice works for a much wider range of people than typing, drawing, or gesture because it is a natural part of human existence. Without a great deal of training, normal human beings can express themselves in a wide variety of domains using voice applications, and thus this breadth of application will be a powerful tool in a ubiquitous environment.

Keywords

Voice Recognition, Speech Synthesis, Digitization ,Acoustic Model, Speech Engine.

1. INTRODUCTION

The system is about to have hand free computer usage using voice recognition and speech synthesis. The idea is to provide an interface which accepts voice as commands and respective action will takes place these may be opening, closing of any application, traversing different drives and folders on the computer, typing of word in editor etc. And each time user gives some command, an audio acknowledgement will be there telling user what action has performed or providing help to the user if command is not recognized. Interface will be user friendly where user will be able to define his own commands and action he want after pronouncing the command.

So here a proposed system will work irrespective of the operating system. Currently it includes mouse as a pointing device and keyboard as typing device. The combination of these two gives a faster way of giving input to the computer and get output equally faster. The latest trend is of using touch sensitive screens to provide input to the computer, but it has its own disadvantages. The proposed system will provide the ultimate 3rd input pattern parallel to mouse and keyboard. This will give a completely new dimension to the computer user interface. It will provide faster means to give multiple inputs at the same time.

Ubiquitous computing (ubicomp) refers to a new generation of computing in which the computer completely permeates the life of the user. In ubiquitous computing, computers become a helpful but invisible force, assisting the user in meeting his or her needs without getting in the way. Ubiquitous computing is a post-desktop model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities. In the course of ordinary activities, someone”using” ubiquitous computing engages many computational devices and systems simultaneously, and may not necessarily even be aware that they are doing so. This model is usually considered advancement from the desktop paradigm. More formally, ubiquitous computing is defined as ”machines that the human environment instead of forcing humans to enter theirs. Similarly in this fashion Voice User Interface tries to fit into the human world by making computer interaction in a more friendly manner. Thus, enabling user to give input in his format rather than following computers.

2. PROPOSED SYSTEM METHODOLOGY

The goal of proposed system is to develop voice applications that can provide intrinsically comfortable, easy-to-use, and efficient to users. As a technology for expression, voice works for a much wider range of people than typing, drawing, or gesture because it is a natural part of human existence. Different audio commands will be used to traverse through the computer, to select the objects displayed on the screen, such as folder or different types of files.

The system will provide the ultimate 3rd input pattern parallel to mouse and keyboard. This will give a completely new dimension to the computer user interface. It will provide faster means to give multiple inputs at the same time. With respect to user each system will be customizable. Each user can generate his own set of commands and use them for different purpose. This will give virtually all control of the computer system to the voice of the user. The VUI in spite of being software with its own interface will be work as a tool for voice conversion. The project work is divided into a number of modules for ease of development. These modules are made by dividing the core activities that are to be performed by the application. Initially at start the idea was as simple as Take

Audio input from User and Convert it to text and perform Operation on Computer.

The Fig.1 shows system architecture of VUI. First of all users activate speech recognition and will give voice input through microphone. Voice will be passed to the filter which will remove unwanted noise. After removing unwanted signals voice will be converted into English words. Converted text will be compared with the default commands in database and finally command is executed by operating system.

- Speech To Text Conversion :

The time required to do the actual Speech to text conversion is user dependent. This activity takes input as user voice and converts it into text. So more the time taken by the user to speak, it increases the total time proportionally. In the command execution part where the words spoken are single or at most in pairs, the conversion time is very less. But for the dictation part where a user will speak many words in continuous way the time for the conversion also increases. But to reduce time if a user speaks too fast then the recognition engine will not understand.

- Command Execution:

This part comes in picture after the spoken word by the user is converted to text and the regarding content is checked in the database. After all this the command to be executed is parsed by the application and is given to the operating system to execute. The execution of the command depends on many external factors of the OS but in general on a stable platform we have checked for the output. The execution time is almost constant irrespective of path depth in the file directory, but if a program requires heavy resources the OS takes care of it.

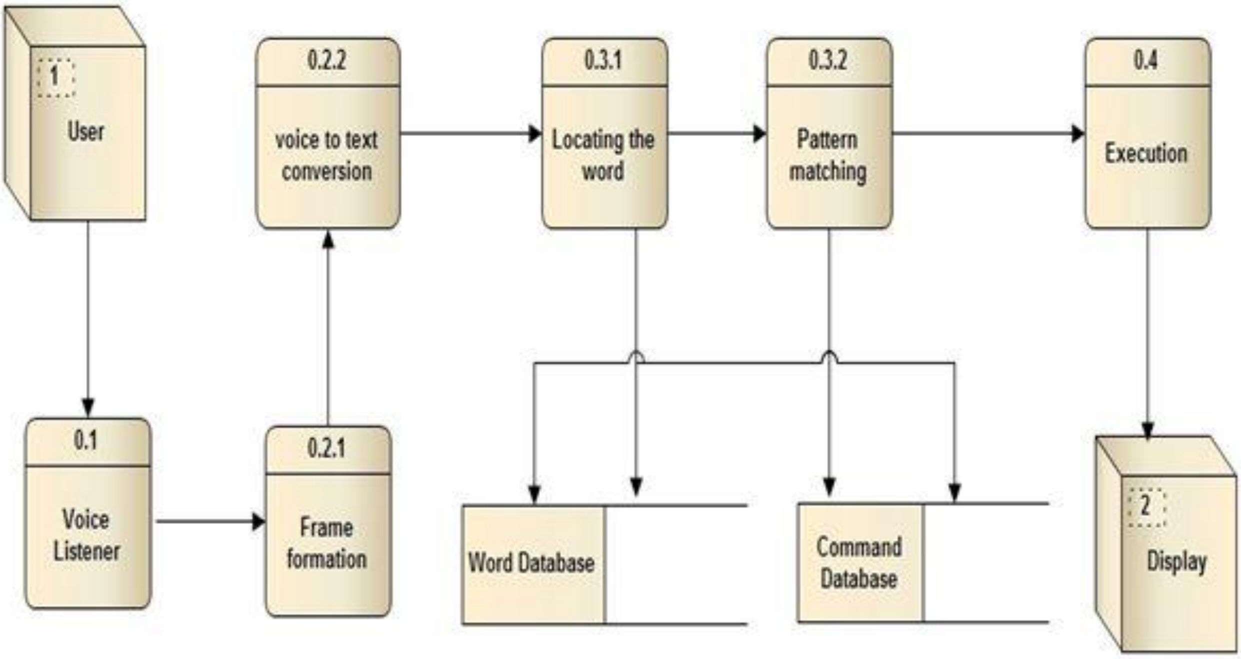


Fig 1: Proposed System Architecture

3. STEPS TO CARRY OUT PROJECT WORK

The project work is divided into a number of modules for ease of development. These modules are made by dividing the core activities that are to be performed by the application. Initially at start the idea was as simple as Take Audio input from User and Convert it to text and perform Operation on Computer. The modules were made in such a way that none of them coincided with each other. The work to be performed by each module is such that there is no dependency on one another.

3.1 Algorithm

This algorithm is general, based on the actual working of the application.

1. Start
2. Enter user credentials
3. If user credentials are correct go to step 3 else go to step 2
4. Activate speech recognition
5. Capture the word spoken by the user
6. Save it and convert it to text
7. If its command execution then go to step 13 else if it is dictation go to step 8
8. Enhance the entered text with respect the word database
9. Check the enhanced text in word database
10. If the word is not present then go to step 8 else go to step 11
11. If multiple words are present for the enhanced word display all of them to the user and wait for user selection else if unique word then type it in the respective area.
12. Wait for next word from the user
13. Enhance the command entered by the user
14. Check for enhanced command in the command database
15. If command is found then go to step 16 else go to step 13
16. Execute the command asked by the user.
17. Display the result of the executed command.
18. Wait for next command from the user or go to step 19

19. Deactivate the speech recognition
20. End

3.2 Searching Database

This is a binary search algorithm that will be implemented for searching the database for words spoken by the user.

```
1. Start
2. The cmd detected command will be searched and database is the database of commands stored in SQLite
3. Calculate mid
4. If(database[mid] == cmd) Return ( action) Else if (database[mid]>cmd) End = mid-1; Goto 3 Else if (database [mid] <cmd) Start=start+1; Goto 3; Else Command not found sorry Return(0)
5. End
```

4. EXPERIMENTAL SETUP

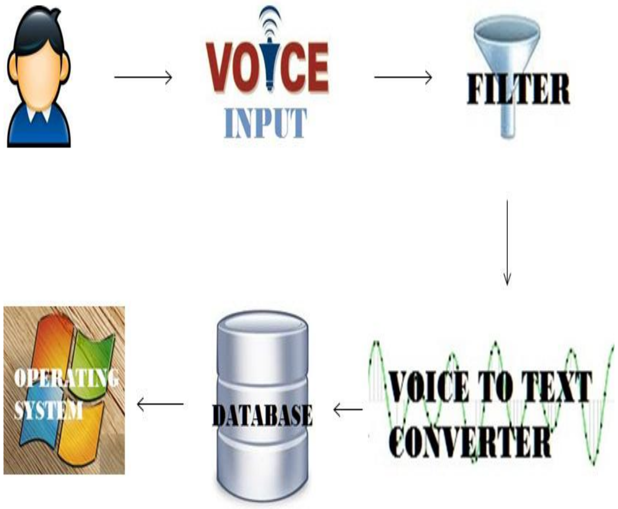


Fig.2: Experimental Setup

The above diagram shows experimental setup of VUI. First of all users activate speech recognition and will give voice input through microphone. Voice will be passed to the filter which will remove unwanted noise. After removing unwanted signals voice will be converted into English words. Converted text will be compared with the default commands in database. And finally command is executed by operating system.

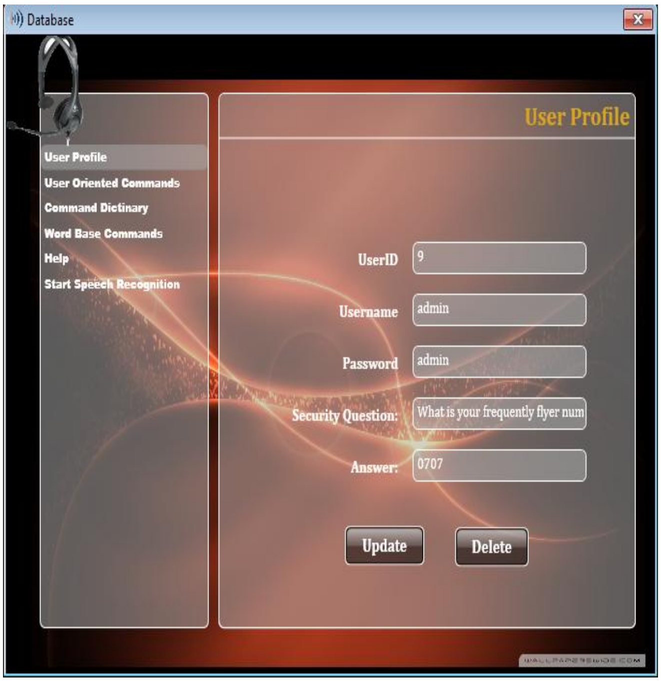
4.1 Healthy Scenarios

- **Computer Traversing:** Authorised user activates speech recognition and said "Open Computer". And My Computer window appear.
- **Typing:** Authorised user activates speech recognition and said "Open Notepad". And said "India is my country. All Indians are my brothers and sisters. I love my country." and text "India is my country. All Indians are my brothers and sisters. I love my country." get typed in notepad at curser position.

4.2 Error Scenarios

- **Computer Traversing:** Authorised user activates speech recognition and said "Open Computer" but system showed message saying "What was that ?" then again user give command "open computer", My Computer window appear.
- **Typing:** Authorised user activates speech recognition and said "Open Notepad". And said "India is my country. All Indians are my brothers and sisters. I love my country." and text "in yet is my country on Indians on my borrow and systems. I luve my NT" get typed in notepad at curser position.

5. RESULTS

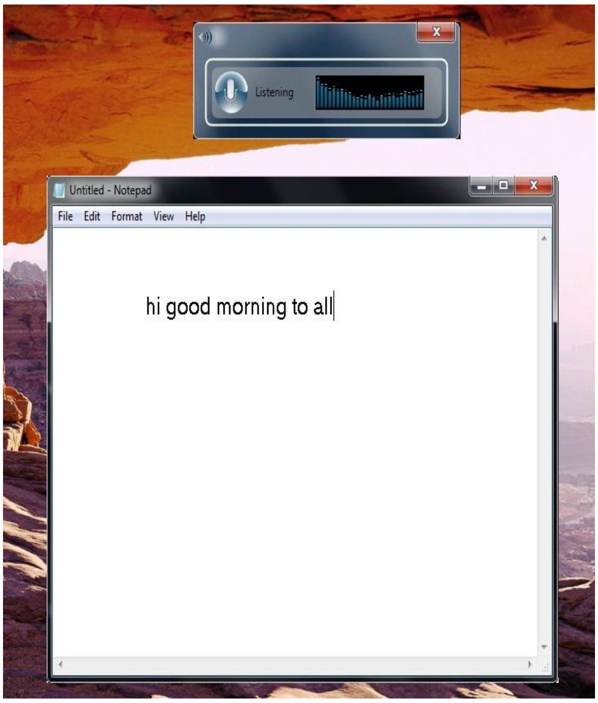


UserID 9
Username admin
Password admin
Security Question: What is your frequently flyer num
Answer: 0707

Fig.3: Home Window

The above Fig.3 shows home window of VUI. It provides different options to user such as:

- **User Profile:** User will be able to update and modify his own information. In the above screenshot user is modifying his information.
- **User Oriented Commands:** It will allow users to add,modify,delete user oriented commands.
- **Microphone Settings:** It will provide options to change your microphone settings.
- **Word Base Commands:** It will allow user to add his own words.
- **Help:** Provide help to operate application.



hi good morning to all

Fig.4: Speech Recognition Window (While writing in notepad)

The Fig.4 shows speech recognition window of VUI. The round button is given for activating speech recognition. Spectrum analyzer is given to show intensity of voice. In above screenshot user is writing to the notepad using voice recognition.

6. ANALYSIS

The basic division of the project has been done into 3 parts from the point of project analysis:

1) Database Access
2) Speech to Text Conversion
3) Command Execution
4) Complete Program Execution

6.1 Database Access

In database access initially the backend to be used was sqlserver. Its performance is very good but required unnecessary overhead. Then embedded database SQLite was used. The method used to access database under SQLite was through the use of SQLite dll package. It satisfied all the needs of the project. But reduction in time constraint was a very necessary factor. So a different technique of using wrapper classes to access the dll directly instead of using .net Framework default SQL engine was decided. The overall requirement of time was reduces by half.

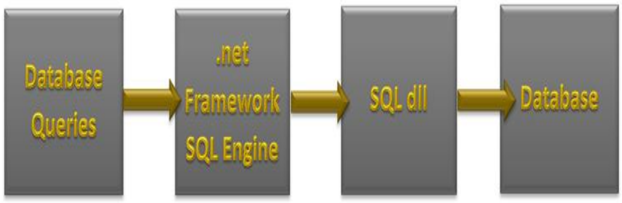


Fig.5: Initial database access

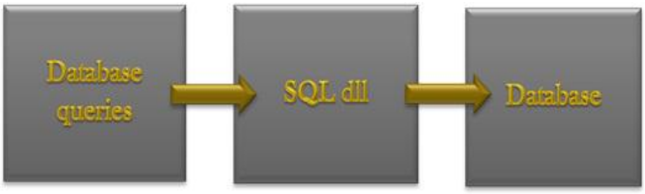


Fig.6: Database access under wrapper classes

Authorised user activates speech recognition and whispers "India is my country". and text "the fifth and have it has nine that he “get typed in notepad at curser position. After whispering the same thing it typed something else.

6.2 Speech To Text Conversion

The time required to do the actual Speech to text conversion is user dependent. This activity takes input as user voice and converts it into text. So more the time taken by the user to speak, it increases the total time proportionally. The command execution part where the words spoken are single or at most in pairs, the conversion time is very less. But for the dictation part where a user will speak many words in continuous way the time for the conversion also increases. But to reduce time if a user speaks too fast then the recognition engine will not understand.

6.3 Command Execution

This part comes in picture after the spoken word by the user is converted to text and the regarding content is checked in the database. After all this the command to be executed is parsed by the application and is given to the operating system to execute. The execution of the command depends on many external factors of the OS but in general on a stable platform we have checked for the output. The execution time is almost constant irrespective of path depth in the file directory, but if a program requires heavy resources the OS takes care of it.

6.4 Complete Program Execution

At a given instance in the lifecycle of the application either the command execution or dictation more will be active. So they won’t interfere in one another.so the actual execution time is just the addition of the required modules time consumption. The project work is compared with a professional product available in market, Dragon naturally speaking, and found out that proposed system works in competition with it, its sometimes slow but on the other hand its low on resources

7. GRAPHS

7.1 Read Single Value from Database

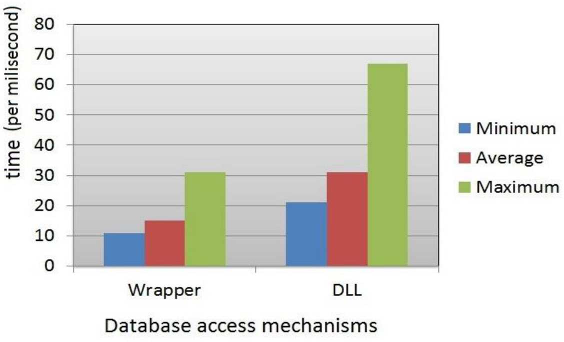


Fig.7: Read Single Value from Database

The above graph shows retrieval of single values into database by using different database access mechanisms. Y axis represents time, whereas X axis represents different database access mechanisms.

7.2 Read Multiple Value from Database

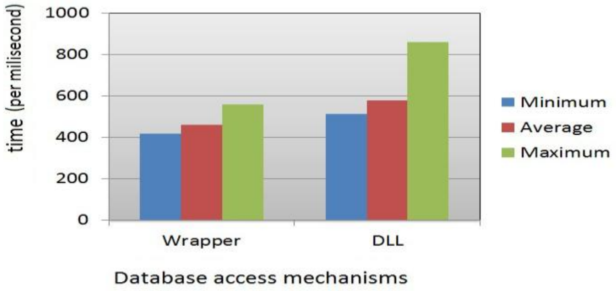


Fig.8: Read Multiple Values from Database

The above graph shows retrieval of multiple values into database by using different database access mechanisms. Y axis represents time, whereas X axis represents different database access mechanisms.

7.3 Insert Single Value into Database

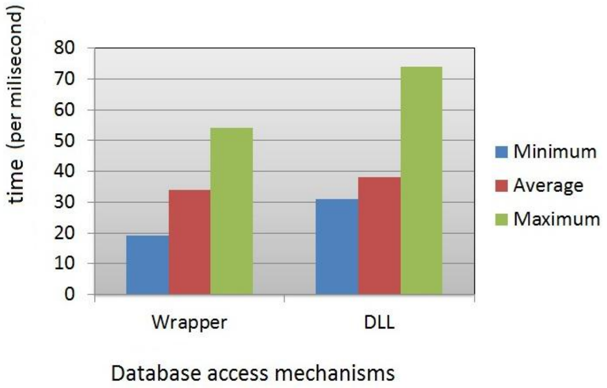


Fig.9: Insert Single Values into Database

The above graph shows insertion of single values into database by using different database access mechanisms. Y axis represents time, whereas X axis represents different database access mechanisms.

7.4 Insert Multiple Values into Database

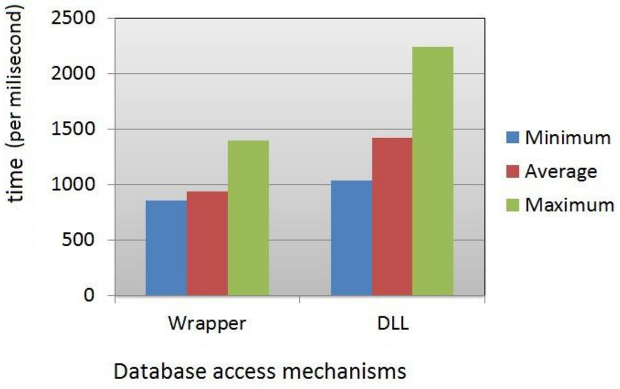


Fig.10: Insertion Multiple Values into Database

The above graph shows insertion of multiple values into database by using different database access mechanisms. Y axis represents time, whereas X axis represents different database access mechanisms.

7.5 Command Execution

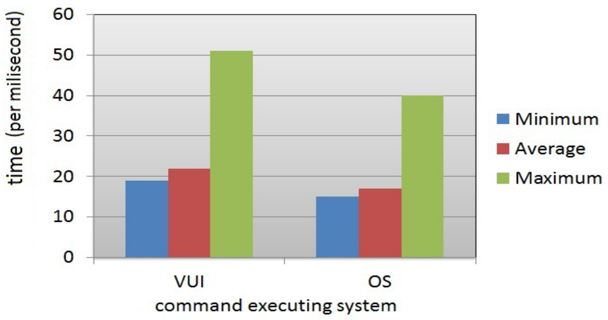


Fig.11: Command Execution

The above graph shows comparison between VUI and other system with respect to command execution. Y axis represents time, whereas X axis represents VUI and other systems.

7.6 Single Word Conversion

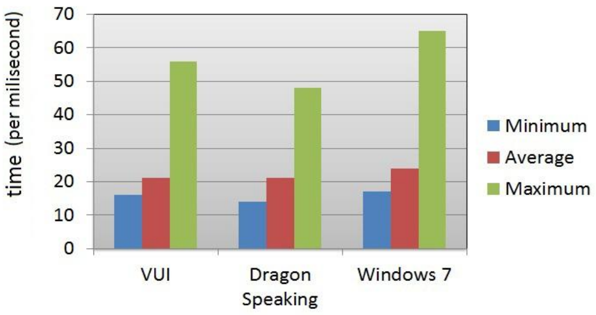


Fig12: Single Word Conversion

The above graph shows comparison between VUI and other system with respect to single word conversion. Y axis represents time, whereas X axis represents VUI and other systems.

7.7 Sentence Conversion

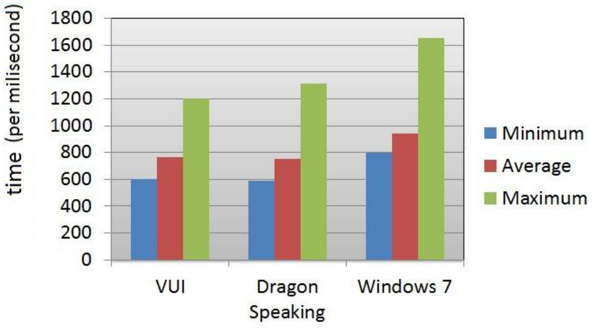


Fig.13: Sentence Conversion

The above graph shows comparison between VUI and other system with respect to multiple word conversion. Y axis represents time, whereas X axis represents VUI and other systems.

8. CONCLUSION

.This project work of speech recognition started with a brief introduction of the technology and its applications in different sectors. The project part of the report was based on software development for speech recognition. At the later stage we discussed different tools for bringing that idea into practical work. After the development of the software finally it was tested and results were discussed, few deficiencies factors were brought in front. After the testing work, advantages of the software were described and suggestions for further enhancement and improvement were discussed.

9. REFERENCES

[1] James R.Evan, Wayne A.Tjoland: Achieving a hand free computer Interface using voice recognition and speech synthesis(2000)

[2] Mukherjee, R.: Text dependent speaker recognition using shifted MFCC (2012)

[3] Robert Keefer, Yan Liu, and Nikolaos Bourbakis: The Development and Evaluation of an Eyes-Free Interaction Model for Mobile Reading Devices (JAN 2013 IEEE)

[4] Vicente P. Minotto, Carlos B. O.Lopes: Audiovisual Voice Activity Detection Based on Microphone Arrays and Color Information (FEB 2013 IEEE)

[5] Xueliang Huo, Hangue Park: A Dual-Mode Human Computer Interface Combining Speech and Tongue Motion for People with Severe Disabilities (NOV 2013 IEEE)

[6] Kwang-Ho Kim, Donghyun Lee: Self -Improvement of Voice Interface with User-input Spoken Query at Early Stage of Commercialization (NOV 2013 IEEE)

[7] A.Esposito, G. Pelosi [8]: Building The Next Generation Of Personal Digital Assistants(2013)

[8] Zhizheng Wu, Xiong Xiao: Synthetic speech detection using temporal modulation Feature(2013)

[9] Tredinnick, R.: Poster: Say it to see it: A speech based immersive model retrieval system(2013)

[10] Vicente P. Minotto, Claudio R.Jung: Simultaneous-Speaker Voice Activity Detection and Localization Using Mid-Fusion of SVM and HMMs (JUNE 2014 IEEE)

[11] Gemmeke, J.F. : The self-taught vocal interface (2014)