EasyChair Preprint

№ 9955

First Steps Towards Speech Controlled Navigation in Web Virtual Reality Environments

Despoina Tsavalou and Vasileios Komianos

April 10, 2023

First Steps Towards Speech Controlled Navigation in Web Virtual Reality Environments

Despoina Tsavalou [1], Vasileios Komianos [1]

[1]*Department of Audio and Visual Arts, Ionian University*

Abstract

In this paper, the first steps of designing and implementing a speech-controlled navigation mechanism for web VR environments are presented. By creating applications that utilize this approach the users can easily navigate in virtual environments and interact with them. Additionally, speech recognition for user input could serve as an efficient approach for disabled users. The mechanism design is presented and brief testing and evaluation are performed in order to assess its effectiveness. The results show that speech commands can be used for navigation in VR environments and provide valuable feedback for future improvements of the proposed approach.

Keywords

speech interaction, web VR, A-Frame, voice input, disabled users, speech recognition APIs

1. Introduction

Navigating Virtual Reality Environments (VREs) is a fundamental task as users entering such environments have to dynamically update their position and explore their spaces. In order to perform navigational tasks, users have to use one of the available navigational methods which are usually dependent on users' ability to move themselves or parts of their bodies. Given that users may not be able to perform such movements, alternative methods can be especially valuable. This paper presents an approach for navigation in Web VR using speech commands.

An important aspect of this work is to explore current technologies in the field of Web Virtual Reality and speech recognition, upon which the proposed approach for navigation in Web VR using speech commands will be implemented. Speech interaction has been proposed as an efficient approach for the interaction of arm disabled users [1,2]. Given the above, a major source of motivation toward this research is the potential limitations that people with disabilities face when interacting with Virtual Reality environments as conventional Virtual Reality navigation approaches require the user to use his hands or other parts of his/her body.

This work focuses on web VREs due to a number of advantages, e.g., the ability to work on all operating systems, the compatibility with modern virtual reality devices, and the fact that such applications can be found via the web and be used without an installation process which provides the potential to reach more users. There is a sufficient number of web VR frameworks that can be used for developing VR applications, e.g., the A-Frame [3], the Vizor Patches [4], and the WebVR-Boilerplate [5].

Speech interaction is already a helpful tool with everyday applications, probably some of the most common are mobile and computer assistants like Siri, Alexa, or Google assistant. Speech recognition

[1]DCAC 2021: 3rd International Conference on Digital Culture & AudioVisual Challenges, Interdisciplinary Creativity in Arts and Technology, Online May 28-29, 2021
EMAIL: despotsabalou@gmail.com (D. Tsavalou); vkomianos@ionio.gr (V. Komianos)
ORCID:0000-0002-1955-6135 (V. Komianos)

for these applications is usually provided by Application Programming Interfaces (APIs), e.g., Google Speech API (also known as web speech API) [6,7], or Vocapia speech to text API [8].

In this work, a demo virtual environment as well as a software library integrating the speech recognition functionality, are developed within the A-Frame framework. For the implementation of the speech recognition functionality, the google speech API is used as it is easily accessible through the web and has a low error rate of 9% [9]. The demo application, called VR-SONIC, is then tested and the results are analyzed in order to provide suggestions for future improvements.

The remainder of this work is organized as follows. Section II provides a brief literature review on works employing speech recognition for interaction mechanisms and instructional purposes. Section III presents the design and the initial implementation of the Speech Controlled Navigation and Interaction mechanism and its functionality. Section IV presents a brief test and the results, while Section V summarizes this work and draws guidelines for future work.

2. Related Work

There is a growing interest in taking advantage of speech recognition for user interaction providing numerous examples and various applications. There are cases where voice input is considered to be an appropriate approach for applications outside the virtual reality domain but due to the nature of the application researchers decided to simulate the application scenarios in VREs and thus are considered to be related and are discussed here. In [10] a car navigation system is introduced, which includes gestural, gaze-based, and speech interaction mechanisms. This navigation system is set up in a virtual reality environment allowing the users to experience an alternative and safe way of driving. This work utilizes the Microsoft speech API in a virtual reality environment developed in Unity. Moreover, in [11] the authors present a voice instructed robot, operating in a virtual reality environment, for law enforcement purposes. In this research, the Julius voice recognition system is used as well as the unity game engine. In this system, the user is requested to press a key, then talk to the robot and the robot responds to a number of commands by making different sounds according to the given command.

Besides the cases that the use of speech input is applied outside the virtual reality domain there are works exploring the capabilities of this technology in the particular domain. In [12] a case study is discussed in order to determine which user interface is preferable for what kind of object interaction in immersive virtual reality. They compare 3 kinds of user interfaces, namely: i) 2D user interfaces, ii) 3D user interfaces, and iii) speech interfaces. They test them with a number of tasks like selection, manipulation, position, rotation, creation, modification, and text input. For the speech interface specifically, they concluded that it performed better than the others but further improvements are needed. In [13] the use of speech input for virtual reality applications is studied by conducting tests to compare various aspects of human-to-machine interaction to human-to-human interaction. The tests show that during human-to-machine speech interaction the participants needed more time to complete a task and that they used a relatively small set of commands consisting of fewer words than the commands given during human-to-human interaction. Moreover, the authors conclude that "*a combination of speech with other input devices might offer users a more flexible and integrated set of interaction tools for VR applications of the future*". In [14] an ongoing effort to develop an interface using input from voice, hand gestures, and eye gaze to interact with information in a virtual environment is presented. For the speech recognition functionality an open source speech recognition library is used (https://snowboy.kitt.ai). It is a key word speech recognition library that runs on raspberry pi hardware and requires to be trained by its user by repeating the keywords to be used a number of times (3). In [15] the use of voice input combined with hand-tracking in order to support positioning, object identification, information mapping, and disambiguation in VREs is explored. The presented application is built with Unity for the Oculus Quest standalone VR headset and uses Microsoft Azure's Cognitive Services for automatic speech recognition.

The contribution of this work is that it is intended to provide speech recognition in web VREs without the need for any special virtual reality equipment and the steps towards this purpose are presented in the following sections.

3. VR-SONIC: Speech Controlled Navigation and Interaction Demo Application

VR-SONIC is intended to allow users to navigate a virtual scene via voice commands and interact with objects in it. The application is aimed at people with disabilities or users who prefer to work through speech recognition. VR-SONIC demo, is a test application developed with A-Frame that incorporates the custom implemented VR-SONIC library which makes use of automatic speech recognition with the Google speech API. A basic goal of the proposed approach is that the provided applications should be easily accessible with the use of a web browser, the applications should be easy-to-use and require as little as possible actions requiring arm movements and that the voice input would be also easy-to-use. Given the aforementioned, A-Frame is selected as it provides the ability to develop web accessible virtual environments, and Google Speech API once integrated in an application is easy-to-use and does not require previous training of the system in order to recognize the commands.

3.1 Prerequisites and User Requirements

VR-SONIC provides an easy-to-use navigation system with simple interaction commands in a virtual reality scene. The user is required to use a personal computer or smartphone in order to access the application. Also, for the application to work properly, the users' microphone must be activated. It is important that the user is able to pronounce the given commands, as speech interaction is the core function of this application. Also, a network connection is needed as the speech recognition is performed by a web API (web Speech API) [4].

For the proper operation of the application, the user is required to follow a series of steps. Firstly, the user has to press the space key (when using a personal computer equipped with a keyboard) in order to enable the microphone. It is stated that despite the fact that arm-controlled operations should be avoided in the presented approach this action is necessary as the user has to provide his/her approval before the browser can use the microphone due to privacy concerns. Once the browser, and the VR-SONIC application, can access the microphone, the user pronounces a command in a short time period. When the command is recognized, the application will update the user’s position in the virtual reality scene according to the given command. That means that the user can either move a pre-set space forward, backward, or turn a pre-set amount left or right, depending on the command they chose.

The current command list integrated into the mechanism is meant to support basic navigation tasks and consists of four commands: i) “move forward”, ii) “move backward”, iii) “turn left” and iv) “turn right”. The small size of the commands list is preferred for this phase of the research in order to provide an easy to debug system while having easily memorized commands. In addition, the use of a limited set of spoken commands is considered to be able to improve the accuracy and speed of a speech recognition system [14].

3.2 User’s position and rotation model

In order to update the user’s position according to the given commands, the mechanism integrates a model defining the user’s position and rotation. Updating a user's position requires knowledge about current position-rotation regarding the world coordinates. The movement is performed on two axes, the x and z, and the y axis is used for the user's rotation so that turning action is allowed (Figure 1). The direction either forward or backward is calculated depending on the user's current rotation. As the distance traveled each time can't be indefinite the step forwards or backward is predefined to be 2 units (meters) and 45 degrees rotation for each time that a turning command is given (Code example 1).

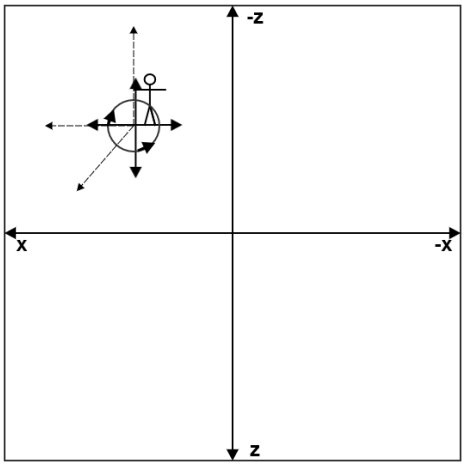

Figure 1: User's position and rotation model

```
if ( command == 'left' || command == 'turn left' )
{
	player.components['look-controls'].yawObject.rotation.y += 0.7853981634;
}

if ( command == 'right' || command == 'turn right' )
{
	player.components['look-controls'].yawObject.rotation.y -= 0.7853981634;
}
```

Code example 1: User's rotation according to the given commands.

4. Testing and Evaluation

In order to evaluate the effectiveness of the proposed solution, an initial test is performed. The user is required to perform basic navigation tasks in order to approach a number of objects in a virtual environment (Figure 2). In each test session, the user is required to approach 3 objects and these tasks are performed with the speech recognition navigation method as well as with the conventional navigation method (mouse & keyboard) for comparison reasons. In both cases, the time required to complete the task is measured. Moreover, in the case of navigation through speech recognition, the number of commands needed in order for the user to reach the objects is recorded and the number of unrecognized commands is counted.

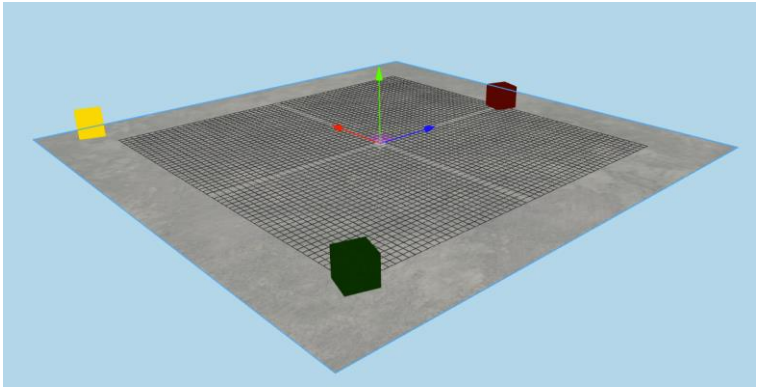

Figure 2: The testing virtual environment

The tests show the time required to complete a navigation task when using the speech commands is significantly longer than the time required compared to the conventional navigation method. Specifically, according to our tests, navigation through speech recognition requires 85,41 seconds on average which is about 20 times more time than the 4,67 seconds that is the average time needed in

order to complete a conventional navigation task (Figure 3). It is also observed that speech commands are not always recognized, and the achieved error rate is 10,13%, which is close to the 9% found by other researchers [7].

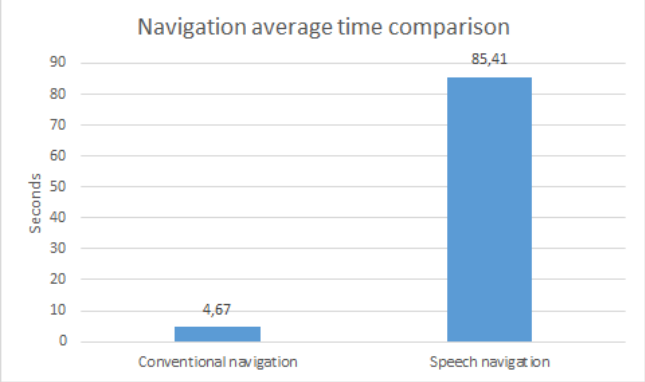

Figure 3: Navigation time comparison

4.1 Discussion

Another observational finding is that the predefined values require a large number of repetitive commands to be given by the user, which increases the time needed in order to perform a navigation task. Performing analysis on the collected data, revealed that the average number of commands in the tested tasks is 15. Also, 2,25 seconds is the average time needed for a speech command to be given by the user. Additionally, the sequence of the recorded commands in each task is analyzed showing that there is an 85% possibility for a user to repeat a move forward/backward command and a 33% possibility to repeat an orientation change move (turn left/right). According to the above, it is calculated that 30% of the total time needed to perform a navigation task is spent in repeating commands (figure 4).

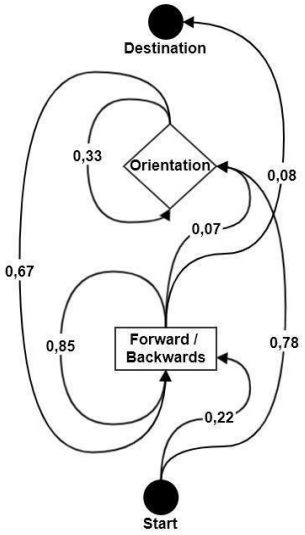

Figure 4: Commands sequence diagram

According to the previous, we consider the improvement of VR-SONIC by introducing commands with parameters. For example, the movement forward or backward is now set to move the controller by 2 units (meters), whereas the number of steps as a parameter could be added so that users do not have to repeat the move commands with such high frequency. Also, the turn command currently performs a 45^{o} rotation, whereas the angle parameter could be integrated in order to provide precise control in rotation.

5. Conclusion

In summary, this work describes the first steps of the design and implementation of a mechanism for providing speech-controlled navigation in web VR environments. The need for easy-to-use methods for navigating VR environments by allowing users not to use their arms or other parts of their bodies is the motivation for this work and provides the ability to interact for arm-disabled users. This first implementation of the proposed approach shows that users can use speech commands to effectively navigate in VR environments. Tests show that there is space for improvements and a number of such additions are briefly discussed and draws the guideline for future work. Moreover, this mechanism can be expanded to include more commands which can be combined with parameters for precise navigation control and for interaction with objects in the VR environments.

Acknowledgements

This work was supported in part by project "Corfu Virtual Exhibition Site for Tourism-Culture-Environment (v-Corfu)," (MIS 5031252), which is partially funded by the European and National Greek Funds (ESPA) under the Regional Operational Programme "Ionian Islands 2014–2020".

References

[1] K. A. Darabkh, L. Haddad, S. Z. Sweidan, M. Hawa, R. Saifan, S. H. Alnabelsi, An Efficient Speech Recognition System for Arm-Disabled Students Based on Isolated Words, Computer Applications in Engineering Education 26 (2018) 285–301.

[2] A. Ferracani, M. Faustino, G. X. Giannini, L. Landucci, A. Del Bimbo, Natural Experiences in Museums Through Virtual Reality and Voice Commands, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1233–1234.

[3] S. Neelakantam, T. Pant, Introduction to A-Frame, in: Learning Web-based Virtual Reality, Springer, 2017, pp. 17–38.

[4] Vizor Patches website, https://patches.vizor.io/, (accessed September 7, 2021).

[5] B. Smus. WebVR Boilerplate. Github. url: https://github.com/borismus/webvr-boilerplate

[6] Google speech API online resource, https://developers.google.com/web/updates/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API, (accessed September 7, 2021).

[7] Web speech API online resource, https://wicg.github.io/speech-api/, (accessed September 7, 2021).

[8] Vocapia speech to text API, website, https://wicg.github.io/speech-api/, (accessed September7, 2021).

[9] V. Këpuska, G. Bohouta, Comparing Speech Recognition Systems (Microsoft API, Google API and CMU Sphinx), Int. J. Eng. Res. Appl 7 (2017) 20–24.

[10] A. Riegler, A. Riener, C. Holzmann, AutoWSD: Virtual Reality Automated Driving Simulator for Rapid HCI Prototyping, in: Proceedings of Mensch und Computer 2019, 2019, pp.853–857.

[11] C. R. Hudson, C. L. Bethel, D. W. Carruth, M. Pleva, S. Ondas, J. Juhar, Implementation of a Speech-Enabled Virtual Reality Training Tool for Human-Robot Interaction, in: 2018World Symposium on Digital Intelligence for Systems and Machines (DISA), IEEE, 2018,pp. 309–314.

[12] D. Hepperle, Y. Weiß, A. Siess, M. Wölfel, 2D, 3D or Speech? A Case Study on Which User Interface is Preferable for what Kind of Object Interaction in Immersive Virtual Reality, Computers & Graphics 82 (2019) 321–331.

[13] A. W. Stedmon, H. Patel, S. C. Sharples, J. R. Wilson, Developing Speech Input for Virtual Reality Applications: A Reality-Based Interaction Approach, International Journal of human-computer studies 69 (2011) 3–8.
[14] J. T. Hansberger, C. Peng, V. Blakely, S. Meacham, L. Cao, N. Diliberti, A Multimodal Interface for Virtual Information Environments, in: International conference on human-computer interaction, Springer, 2019, pp. 59–70.
[15] J. Sin, C. Munteanu, Let's Go There: Combining Voice and Pointing in VR, in: Proceedings of the 2nd Conference on Conversational User Interfaces, 2020, pp. 1–3.

Information about the authors

Despoina Tsavalou is an undergraduate student in the Department of Audio Visual Arts, of the Ionian University. During her studies she deals with digital audiovisual arts, new technologies and media (internet, multimedia) and applications with human-centered orientation. Recently she has been doing research in the field of digital interactions and their practical applications.

Vasileios Komianos is a faculty member at Dept. of Audio and Visual Arts, Ionian University, Greece, teaching courses related to Virtual/Augmented/Mixed Reality, video games and interactive multimedia. His research interests are mostly focused on Mixed Reality (MR) systems, on user interaction and user interfaces in MR systems and applications as well as on approaches for artistic expression and cultural communication. He has work experience on designing audiovisual content and installations in the cultural heritage sector, and his works are hosted or have been hosted in permanent and temporary exhibitions as well as in art festivals.