MICROSOFT SPEECH RECOGNITION AS A NEW WAY OF COMMUNICATION IN AUGMENTED REALITY ENVIRONMENT

NGU WEN WOO

A thesis submitted in fulfilment of the requirements for the Degree of Master of Science (Cognitive Science)

Faculty of Cognitive Sciences and Human Development

UNIVERSITI MALAYSIA SARAWAK

2013
AUTHOR DECLARATION

I hereby declare that this dissertation entitled “Microsoft Speech Recognition as a New Way of Communication in Augmented Reality Environment” is the result of my own research excepted as cited in the references. The thesis is not concurrently submitted in candidature of any other degree.

______________________
Ngu Wen Woo

Department of Cognitive Sciences

Faculty of Cognitive Sciences and Human Development

Universiti Malaysia Sarawak
ACKNOWLEDGEMENT

I would like to dedicate my special thanks to my supervisor, Dr. Ng Giap Weng for his guidance, comment and support throughout the project.
TABLE OF CONTENTS

AUTHOR DECLARATION ii
ACKNOWLEDGEMENT iii
TABLE OF CONTENTS iv
LIST OF FIGURES x
LIST OF TABLES xiv
ABSTRACT xv

CHAPTER 1 INTRODUCTION 1

1.0 Overview 1
1.1 Background of Literature 2
1.2 Problem Statement 5
1.3 Objective 6
   1.3.1 General Objective 6
   1.3.2 Specific Objectives 7
1.4 Importance of Research 7
1.5 Research Scope 8
1.6 Research Value 9
1.7 Significance of Research 10
1.8 Limitations 10
1.9 Definition of Terms 10
1.10 Summary 11
### 2.12 Summary

<table>
<thead>
<tr>
<th>3.0 Overview</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Research Framework</td>
<td>38</td>
</tr>
<tr>
<td>3.2 System Development</td>
<td>40</td>
</tr>
<tr>
<td>3.2.1 System Specification</td>
<td>43</td>
</tr>
<tr>
<td>3.2.1.1 Hardware</td>
<td>43</td>
</tr>
<tr>
<td>3.2.1.2 Software</td>
<td>45</td>
</tr>
<tr>
<td>3.2.1.2.1 3Ds Max</td>
<td>45</td>
</tr>
<tr>
<td>3.2.1.2.2 VRML</td>
<td>45</td>
</tr>
<tr>
<td>3.2.1.2.3 ARToolkit</td>
<td>46</td>
</tr>
<tr>
<td>3.2.1.2.4 ARUT</td>
<td>47</td>
</tr>
<tr>
<td>3.2.1.2.5 Microsoft Visual Studio C++ 2008 Express</td>
<td>48</td>
</tr>
<tr>
<td>3.2.1.2.6 Microsoft Direct X</td>
<td>38</td>
</tr>
<tr>
<td>3.2.1.2.7 OpenGL</td>
<td>49</td>
</tr>
<tr>
<td>3.2.1.2.8 Microsoft Speech 5.1 SDK</td>
<td>49</td>
</tr>
<tr>
<td>3.2.1.3 System Requirements</td>
<td>50</td>
</tr>
<tr>
<td>3.2.2 Object Collection</td>
<td>50</td>
</tr>
<tr>
<td>3.2.3 System Design</td>
<td>51</td>
</tr>
<tr>
<td>3.2.4 Coding</td>
<td>53</td>
</tr>
<tr>
<td>3.2.5 Testing</td>
<td>53</td>
</tr>
</tbody>
</table>
3.1.6 Integration

3.3 Pilot Test

3.4 Evaluation

3.5 Summary

CHAPTER 4 INTEGRATION AND DEVELOPMENT

4.0. Overview

4.1 Speech Recognition SDK 5.1

4.1.1 Speech Recognition Basic

4.1.2 Speech Recognition Flow

4.1.3 Speaker Profile

4.1.4 Context Free Grammar

4.1.4.1 Grammar Format

4.1.5 Retrieve Recognition Information

4.1.6 Command and Control

4.2 Interaction

4.3 Marker Detection

4.4 Sound Effect

4.5 Video File

4.6 Model Selection

4.6.1 3D Modelling

4.6.1.1 Line

4.6.1.2 Mesh
5.5.1 Demographic Data 116
5.5.2 Functionality and Performance 118
5.5.3 Acceptance 122
5.6 Discussion 126
5.7 Strengths 129
5.8 Limitation 130
5.9 Summary 131

CHAPTER 6 CONCLUSION 133
6.0 Overview 133
6.1 Contribution 133
6.2 Recommendation for Future Study 135
6.3 Summary 136

REFERENCES 139

APPENDICES 149
Appendix A Consent Form 149
Appendix B User Background Information 150
Appendix C Functionality and Performance User Tasks 151
Appendix D System Effectiveness User Feedback 153
Appendix E Certificates of Achievement 155
# LIST OF FIGURES

<p>| Figure 2.1 | Milgram Reality Virtuality Continuum | 14 |
| Figure 2.2 | Example of 2D square markers | 16 |
| Figure 2.3 | Example of ARTag markers | 17 |
| Figure 2.4 | HMM prediction for “tomato” | 19 |
| Figure 2.5 | The use of mousegrid with nine zones number | 28 |
| Figure 2.6 | Nuance product “Dragon Naturally Speaking 10” | 30 |
| Figure 3.1 | Overall Research Framework | 38 |
| Figure 3.2 | System Development Life Cycle | 42 |
| Figure 3.3 | ARToolkit illustration | 47 |
| Figure 3.4 | System flow | 51 |
| Figure 4.1 | ISpRecognizer | 63 |
| Figure 4.2 | ISpRecoContext::CreateGrammar | 64 |
| Figure 4.3 | ‘ISpRecoGrammar’ | 65 |
| Figure 4.4 | Control Flow Diagram | 66 |
| Figure 4.5 | Code snippet of grammar file loading | 68 |
| Figure 4.6 | XML grammar files | 69 |
| Figure 4.7 | Code snippet for recognition retrieving | 70 |
| Figure 4.8 | wcstombs | 71 |
| Figure 4.9 | std::string and std::compare use in words comparison | 72 |
| Figure 4.10 | Snippet of code to get user input | 73 |
| Figure 4.11 | Marker Detection | 74 |
| Figure 4.12 | Converting ARToolkit matrix to OpenGL matrix | 75 |
| Figure 4.13 | Snippet of Windows API for sound effect | 76 |
| Figure 4.14 | Snippet of code for video loading, playing, pause and stop code | 77 |
| Figure 4.15 | A number “one” model being developed using ‘line’ | 79 |
| Figure 4.16 | 3D solid object of ‘one’ | 79 |
| Figure 4.17 | A cat model been developed using ‘editable mesh’ | 80 |
| Figure 4.18 | ‘Material editors’ interface for texture mapping | 81 |
| Figure 4.19 | Models with applied material | 82 |
| Figure 4.20 | Model animation | 83 |
| Figure 4.21 | Example of the algorithm in VrmlPad | 84 |
| Figure 4.22 | Coding line for the display of the first ‘.wrl’ file on call | 85 |
| Figure 4.23 | System setting | 85 |
| Figure 4.24 | Example of the system output | 86 |
| Figure 4.25 | Snippet of code for detecting the user speech input | 87 |
| Figure 4.26 | Code snippet for word comparing and text output | 88 |
| Figure 4.27 | Example of the output in text form | 88 |
| Figure 4.28 | The display of ‘.wrl’ file | 89 |
| Figure 4.29 | ‘.wrl’ files listed on the system setting | 90 |</p>
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.30</td>
<td>Code snippet for activating category content</td>
<td>90</td>
</tr>
<tr>
<td>4.31</td>
<td>Code snippet for controlling the 3D output on one of the category content</td>
<td>91</td>
</tr>
<tr>
<td>4.32</td>
<td>Number category 3D objects</td>
<td>92</td>
</tr>
<tr>
<td>4.33</td>
<td>Animal category 3D objects</td>
<td>92</td>
</tr>
<tr>
<td>4.34</td>
<td>Insect category 3D objects</td>
<td>93</td>
</tr>
<tr>
<td>4.35</td>
<td>An example of AR system</td>
<td>93</td>
</tr>
<tr>
<td>4.36</td>
<td>Code snippet for displaying an image</td>
<td>94</td>
</tr>
<tr>
<td>4.37</td>
<td>*.bat file content</td>
<td>95</td>
</tr>
<tr>
<td>4.38</td>
<td>Different image shown based on the different speech recognized</td>
<td>95</td>
</tr>
<tr>
<td>4.39</td>
<td>Coding line for the calling of the song genre</td>
<td>96</td>
</tr>
<tr>
<td>4.40</td>
<td>Code snippet for playing a song function</td>
<td>97</td>
</tr>
<tr>
<td>4.41</td>
<td>3D text displayed based on the song name chosen</td>
<td>98</td>
</tr>
<tr>
<td>4.42</td>
<td>Snippet of code for video loading and functions</td>
<td>98</td>
</tr>
<tr>
<td>4.43</td>
<td>Snippet of code for video displaying</td>
<td>99</td>
</tr>
<tr>
<td>4.44</td>
<td>Snippet of code to manipulate video using SR system</td>
<td>99</td>
</tr>
<tr>
<td>4.45</td>
<td>A video played in the AR environment</td>
<td>100</td>
</tr>
<tr>
<td>5.1</td>
<td>Screenshot of fiduciary marker</td>
<td>102</td>
</tr>
<tr>
<td>5.2</td>
<td>Screenshots of registration</td>
<td>103</td>
</tr>
<tr>
<td>5.3</td>
<td>Screenshots of AR enabled speech application in text form</td>
<td>105</td>
</tr>
<tr>
<td>5.4</td>
<td>Screenshots of SR enabled AR 3D objects system</td>
<td>105</td>
</tr>
</tbody>
</table>
Figure 5.5  Cat ‘meow’  107
Figure 5.6  Dog ‘bark’  107
Figure 5.7  Image displayed based on the user defined spoken word  108
Figure 5.8  3D text indicates which song was played  108
Figure 5.9  Screenshot of a video on top of the marker  109
Figure 5.10  Marker that has been moved  111
Figure 5.11  Block marker  112
Figure 5.12  Different camera height  112
Figure 5.13  Different viewing point  113
Figure 5.14  Different lightning condition  114
Figure 5.15  The system mainly make “no recognition” on high level of noise  115
Figure 5.16  Multimedia (text, 3D objects, images and video) application evaluation  120
Figure 5.17  Video in AR application evaluation  121
Figure 5.18  Means and standard deviation of the system evaluation on overall AR application  124
Figure 5.19  Dog stand when ‘stand’ spoken  127
Figure 5.20  Dog lye when ‘lye’ spoken  127
Figure 5.21  Dog produce sound when ‘sound’ spoken  127
Figure 5.22  Example of animation  128
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Differentiation between different types of SR</td>
<td>23</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Hardware Needed</td>
<td>44</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Statement for multimedia (text, 3D objects, images and music) application</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Statements for video in AR application</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Statements for user feedback</td>
<td>56</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Reliability statistic results</td>
<td>57</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Demographic data of the participants in SR as Input Device in AR Environment</td>
<td>117</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>The results of the evaluation on the system functionality and performance of the AR application</td>
<td>119</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Means and standard deviation of the system evaluation on overall AR application</td>
<td>123</td>
</tr>
</tbody>
</table>
The advancement of technology in computers has given rise to more sophisticated and smarter computers. This has spurred on the birth of Augmented Reality (AR) technology. AR enables users to view virtual 3D model in the real world environment in real-time. Mostly, we depend on traditional input devices such as the keyboard, mouse or joystick to interact within the AR environment. Thus, it is crucial to come out with sophisticated type way of communication to suit the current technology. Speech Recognition (SR) can be used to replace traditional input devices. SR enables users to communicate within the AR environment by using spoken words. SR was integrated in the AR environment as a way of communication. Microsoft Speech SDK 5.1 is used to build the necessary SR to be integrated in the AR environment. Users can continuously speak with the computer to execute certain control or commands. 3D models are created and rendered in a Virtual Reality (VR) environment to correspond with the input given. Furthermore, an evaluation was conducted to test the system functionality of SR for interaction with the virtual object in AR system. The result of the evaluation reinforces the potential of SR as a command and control in the AR system.
ABSTRAK

PENGENALAN UCAPAN MICROSOFT SEBAGAI AGEN KOMUNIKASI BARU DALAM PERSEKITARAN REALITI AUGMENTASI

Kepesatan pembangunan komputer dari segi teknologinya telah memberi kebangkitan kepada teknologi komputer yang lebih canggih dan pintar. Ini telah memberi kelahiran kepada Realiti Augmentasi (RA). RA membolehkan pengguna untuk melihat model 3D dalam persekitaran maya dan persekitaran dunia sebenar dalam masa nyata. Kebanyakan kita bergantung kepada alat komunikasi tradisional seperti papan kekunci, tetikus atau kayu ria untuk berinteraksi dalam persekitaran RA. Oleh itu, agen komunikasi yang lebih canggih perlu diperkenalkan untuk dipadankan dengan teknologi terkini. Pengecaman Ucapan (PU) boleh digunakan untuk menggantikan peranti input tradisional. PU membolehkan pengguna untuk berinteraksi dalam persekitaran RA hanya dengan menggunakan sebutan perkataan. PU telah disepadukan sebagai agen komunikasi dalam persekitaran RA. Untuk tujuan ini, PU telah dibina dengan menggunakan Micrsoft Speech SDK 5.1 dan disepadukan dalam persekitaran RA. Pengguna boleh terus bertutur dengan komputer untuk melaksanakan kawalan dan arahan tertentu. Model-model 3D dicipta dan dipamerkan dalam persekitaran realiti maya sebagai padanan dengan input yang diberikan. Di samping itu, penilaian ke atas fungsi sistem SR sebagai interaksi dengan objek maya dalam sistem RA telah dijalankan. Keputusan penilaian menunjukkan potensi PU sebagai kawalan dan arahan di dalam sistem RA.
1.0 Overview

Augmented Reality (AR) is a technology that enable user to render two dimensional (2D) and three dimensional (3D) virtual objects in the AR environment. This technology requires a video capture device in order for the users to capture real time environment and render the AR environment. When the rendering process is on-going, user can either use the mouse or keyboard to interact with the virtual objects.

This research studied on the use of other alternative way of communication to interact with the AR environment. Speech Recognition (SR) is one of the communication methods to enable the system to perform the necessary output which in turn will generate 3D computer graphics in the real environment.
AR also emphasises on the interaction between the users and the virtual objects to perform real-world tasks. As such, the natural speech is used as a way of communication to produce a natural interaction for the users. SR technology is integrated for the execution of natural interaction. This provides an AR system that allows users to interact with the virtual objects based on vocally produced sound.

Chapter 1 discusses the introduction of AR technology and the background of this research. Furthermore, it moves on to the problem statements, general objectives, specific objectives and definition of terms in this research. This chapter also contains the limitation, scope of the projects and significance of the study.

1.1 Background of Literature

The advancement of computers has led on to more sophisticated and smarter computers. As a result, the Virtual Reality (VR) was introduced into the field of computer sciences. VR is a user interface technology that allows user to view the artificial environment in such way that user think and acknowledge it as a real environment (Margaret, 2009). Besides, it enables users to interact with the system in real-time response and immerse in the virtual environment.

From the VR technology, Augmented Reality (AR) has been introduced in order to better interact with the computer effectively. AR is a growing area and also one of the variations in the VR research. There are three key characteristics of AR system have
been identified (Billinghurst, Schmalstieg, & Langlotz, 2011). First is the mixing of the virtual objects with the real world environment. The second is the registration of three dimensional digital data. The virtual objects are superimposed upon in the real world like if they are part of the scene. Lastly, users are able to interact with the virtual objects in real time. It will realize a common goal of an AR system in which users won’t be able to differentiate between the real world and the virtual augmentation of it.

AR has received increasing interest and attention for the last few years. The VR and AR technology offer a possibility for the digitization of the objects in the real world which is known as virtual objects. The AR term has been used to enable users to view virtual 3D objects in the real-world environment in real-time. It is the combination of the real scene being viewed and a computer generated virtual scene. It will enable the users to have a sense of being in the real world and at the same time enhance the user perception of the surrounding world.

Speech Recognition (SR) is also associated with the technology that was born through the advancement in the technological arena which has been presiding for several decades. It has undergone many improvement and achievement throughout the year in becoming a more sophisticated version of SR technology. SR, which is also known as Automatic Speech Recognition (ASR) is able to convert spoken words to text. A computer can now identify converted speech, the spoken words through the microphone using the SR system (John, 2003). This technology can provide a broader
solution in speech recognition in which user can just direct speak to the machine to execute certain task.

The SR technology consists of an SR engine that can determine the performance of the SR process. The primary function of the SR engine is to process the spoken words and translate it into text that is comprehensible by an application (Kimberlee, 2001). In the application, the SR engine functions in two ways. First, the application can interpret the result of the recognition as a command. Furthermore, the application is able to handle the recognized spoken word simply as a text which is considered a dictation. This function enhances the process in recognizing spoken words especially in speech translation.

This project is conducted with the aim of integrating speech technology in AR environment. The application integrates the SR engine in the AR system as a new method of human computer interaction especially in the AR environment. It enhances a new experience since SR allows users to provide speech input the same way we enter inputs through the keyboard and clicking on a mouse (Kimberlee, 2001). Through this integration, it will create a new way of wireless communication between users and at the same time be fully immersive in AR environment. Since the system is able to recognize the users spoken words, they are able to know they have mispronounced when the system could not recognize their speech.
1.2 Problem Statement

As we all know, AR can also serve as the medium of communication between users and 3D objects. Generally, when users view the Virtual Reality (VR) world using the AR technology, they will communicate with the computer using the mouse or keyboard as the only way of communication. There will be more physical contact since users need to place their hands on the mouse or keyboard in order to get the desired action when viewing the VR world. Navigation in the 3D graphic environments can be unnatural when using conventional mouse (Goose, Sudarsky, Zhang, & Navab, 2003). It eventually makes the AR environment less immersive for the user.

Besides, the traditional input devices such as mouse and keyboard might be troublesome for users who prefers their hands to be free when they entering the AR environment. This kind of situation can be avoided if a user wears the head mounted display. Unfortunately, the head mounted display is expensive to be owned by most of the people. Since the keyboard and mouse is synonym to a method of communication, the absence of the keyboard and mouse will halt us from communicating with the computer. Without the alternative method of communication, it is hard for users to feel as if they are in the immersive world. Therefore, there is a need for an effective solution to solve the arising problem.

This research is to design and develop a system that implement Microsoft SR in the AR environment as a way of communication. This system will involve less physical
contact and there will also be more interesting when using the AR technology. However, the new technology also brings about new problems, challenges and the potential for massive mistakes even though there is great potential (Norman, 2010). It need quite some time for the user to extract the best potential out of it.

The implementation of SR technology requires user to speak with the computer for the interaction to take place. Unlike a multimodal which uses the combination of two or more method of interaction, SR is an independent way of communication that relies solely on speech input. Therefore, there is a need to develop a SR system that is able to capture the speech effectively. This research involves developing a SR system as the main way of communication for implementation in the AR environment.

1.3 Objective

The objective is divided into two categories, which are the general objective, and the specific objectives.

1.3.1 General Objective

The general objective of this study is to integrate Microsoft SR with AR technology and fully utilize the SR technology in building alternative way of communication. The interface is developed using AR technology to make immersive environment.
With the integration, it will create a more realistic environment with the flexible communication method.

1.3.2 Specific Objectives

Specifically, the objectives of this study are:

a) To integrate Microsoft Speech Recognition as a way of communication in Augmented Reality Environment.

b) To use speech as command to control the multimedia features (text, 3D objects, images, music and video) in the Augmented Reality environment.

c) To test and evaluate speech as a command to control the multimedia functionalities in Augmented Reality environment.

1.4 Importance of Research

The development of AR has been dispersed into many fields of application. AR technology is an emerging technology which enables user to perceive the VR in a real world environment. Integration of the SR technology has the capability of enhancing the method of communication by using merely spoken words. The SR technology provides broader solutions that can recognize speech without targeting a particular speaker (Wadhwani, Kolhe, & Dekate, 2011).
The application that is going to be developed enables the usage of SR technology as a way of communication. When running the application, related spoken words can be used as command and control to display related 3D objects or models.

Users are allowed to interact with the VR environment in a more interactive and exciting manner. Since the SR technology is used, user can freely speak related words or sentences to view the desired objects in the application. The user can view the animated 3D models and feel immersed at the same time.

This kind of method can help to attract more users to the particular speech integrated application. Users will be keener to view the virtual objects, characters and scene because this method is able to build up the interest of the user. Apart from that, they can develop a more futuristic sense when using this application.

1.5 Research Scope

In this research, the models and objects will be digitized. Through this, it can be displayed in real scene by using the AR technology. SR technology is also integrated as the way of communication in the AR environment. The interaction with the AR system is then executed with the use of SR technology.

The predefined words are created first to act as the command in executing the certain task or controlling the virtual objects in the AR environment. The conventional use of the keyboard and mouse eventually can be neglected. It is hope that it can enhance