Bachelor's thesis

Degree programme in Information and Communications Technology

2021

Chung Pham

USER EXPERIENCE ON SPEECH RECOGNITION IN SERIOUS GAMES

– Case study: ExerGo

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Degree programme in Information and Communications Technology

2021 | 39 of pages, 13 pages in appendices

Chung Pham

USER EXPERIENCE ON SPEECH RECOGNITION IN SERIOUS GAMES

– Case study: ExerGo

The thesis aimed to investigate the advantages and disadvantages of implementing speech recognition in a serious game. Moreover, it purposed to measure factors that could affect the user experience while users were playing a serious game, ExerGo using speech recognition. The thesis also aimed to make some recommendations that could be taken into account by the developers of the case study game using.

The theoretical sections focus on the user experience, serious games, speech recognition technology, and research method. The empirical part was carried out by using two questionnaires. The first questionnaire purposed to explore how people think about serious games and speech recognition. The second questionnaire aimed to discover the players' opinions on speech recognition in the case study game. The first questionnaire was administered to 98 participants and there were 22 testers who tested the ExerGo game.

The results illustrate that implementing speech recognition in serious games has some advantages and has a good impact on user experience. However, the speech recognition caused some dissatisfaction to the participants.

KEYWORDS:

User experience, speech recognition, serious games, AI, artificial intelligence, IBM Watson

CONTENTS

FIGURES

APPENDICES

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AI	Artificial intelligence
API	Application programming interface
AR	Augmented reality
BEE	Business Ecosystems in Effective Exergaming
FIT	Futuristic Interactive Technologies
ISO	International Organization for Standardization
OS	Operating system
PC	Personal computer
SDK	Software development kit
STT	Speech-to-text
TTS	Text-to-speech
UI	User interface
URL	Uniform resource locator
US	United State
UX	User experience
VR	Virtual reality

1 INTRODUCTION

Humans communicate with each other in many ways such as speech, hand gestures, facial expressions. However, speech is considered the most important means that humans use, as it facilitates communication and it is the most widely used means of communication among speakers. Since the Digital Revolution began in the middle of the 20th century, many important inventions were created: the internet, operating systems, and millions of applications. In order to interact with those applications, users must use some traditional input devices such as a mouse, keyboard. These types of input, in some applications, are not so convenient, especially when users are moving or driving or are disabled people. Therefore, speech recognition brings an opportunity for scientists to develop a new way to communicate and exchange information with a machine without physical interaction. Nowadays, speech recognition is very popular in our lives. People can start a car just by saying few words, and by saying “Hey Siri”, iPhone users can start an application or ask any question they want to know. Speech recognition really makes life more comfortable.

This thesis was commissioned by by Futuristic Interactive Technologies (FIT), a research group at Turku University of Applied Sciences. They needed to implement speech recognition into a game called ExerGo and wanted to know how speech recognition affects user experience. The ExerGo is a health mobile application that aims to motivate users to run to improve their health. While running, it is difficult and not safe for users to monitor their progress by focusing their eyes on their phones. Therefore, designers decided to use speech recognition as a voice-user interface to interact with the application. The speech recognition technology was chosen in this project was from IBM because IBM provides an software development kit (SDK) for Unity, the game engine which was used to develop ExerGo, and its effectiveness. The author of this thesis was a game programmer student at Turku University of Applied Sciences and was doing an internship in Turku Game Lab at the time the task was assigned to him.

The purpose of this thesis is to answer the following question: How does speech recognition affect user experience in serious games? This question can be divided into the following sub questions:

- What do users think of speech recognition?

- What factors could improve the user experience in serious games?

- What are the benefits of implementing it to such projects?

- What are the problems users have when they test the game?

The thesis has three objectives:

- To discover participants' opinion about speech recognition.

- To explore factors that influence user experience while users are playing a serious game using the speech recognition.

- To investigate the advantages and disadvantages of implementing speech recognition in the game.

To help readers can follow and have a better understanding of this thesis, the thesis is divided into 6 chapters. Chapter 1 presents the general context of the thesis's topic including the research question, the reasons, and the objectives of the thesis. Chapter 2 introduces key concepts of the thesis. These concepts include user experience, user interface, usability, and serious games. The methodology is also discussed in this chapter. Chapter 3 presents the background information of artificial intelligence and the speech recognition technology and how speech recognition applied in games. Moreover, the latest speech recognition technologies in some companies are compared. Chapter 4 introduces the ExerGo game and explains the steps for implementing speech recognition into the ExerGo game. Chapter 5 presents the key research findings. The final chapter concludes and discusses the limitations of the thesis and suggestions for further research.

2 THEORETICAL BACKGROUND

This chapter will describe some concepts related to the topic. Thus, it will provide basic knowledge for readers to understand what are user experience and serious games. In addition, it will discuss the research methods used in this thesis.

2.1 User Experience

According to the International Organization for Standardization (ISO), user experience (UX) by definition is a “person’s perceptions and responses resulting from the use and anticipated use of a product, system or service”. More simply, UX is how a person feels about things that he or she is interacting with. Those reactions not only happen while a user using or playing something, but they also occur before and after the action. Furthermore, UX is a result of many factors including the user’s opinion about the brand, how the system functions. The user’s previous experiences, skill, behavior and personality also affect the UX (ISO, 2019.)

UX is very important for a product or a service as the final goal of developing a product is to attract customers. In terms of business, a good UX will increase the return on investment. If a new user was impressed while testing a product or service, the user could engage with the organization which made the product or service. Moreover, feelings of satisfaction while playing a game or using a service could increase the loyalty of a customer. This means that it encourages the customer to use the product more frequently. Finally, UX is a part of the development process. In the real world, most products need to be tested carefully before the final release. Testing allows the design and development teams to evaluate their work and to make changes if testers do not feel comfortable with the product. Testing many times with real users guarantees the final product will make the customers happy (Cameron, 2020.)

2.1.1 Differences between User Experience, Usability and User Interface

There seems to be some confusion over the concepts of user experience, user interface, and usability when a product or service is used. However, there are many differences between those terms.

ISO defines usability as the effectiveness, efficiency, and satisfaction of completing particular targets by specified individuals using a product, system, or service in a

specified situation (ISO, 2019). Jakob Nielsen who is considered as “the king of usability” defined usability by 5 quality components: learnability, efficiency, memorability, errors, satisfaction (Web Designer Depot, 2009; Nielsen, 2012). Learnability means how quickly users become familiar with the product for the first time they use it. Efficiency means how fast users complete the given tasks. Memorability is the ability to reestablish proficiency from users when they use the product again. Errors refer to how many errors users make and their ability to recover after making the errors. Finally, satisfaction is how pleasant it is for the user to use the product (Nielsen, 2012.)

From the definitions of user experience and usability, it is clear that UX will answer the question of how satisfied a user is while usability refers to how easy to use a product or a service. In fact, usability is one necessary part of user experience. Therefore, user experience is more important than usability in terms of making successful products. Figure 1 shows 7 factors that make user experience including usability (Interaction Design Foundation, 2020.)

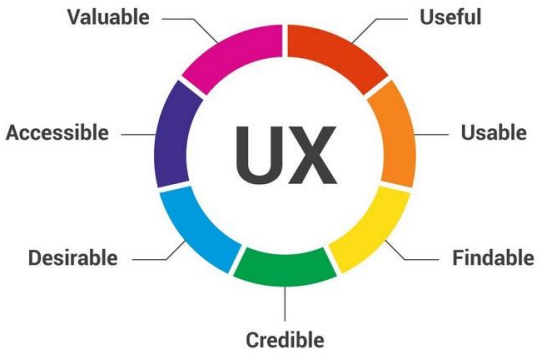

Figure 1. User Experience Factors (Interaction Design Foundation, 2020).

The term user interface (UI) refers to both the physical and the digital interaction between a human or operator and a device or equipment (Luchs, 2016). The key difference between UX and UI is that UI is all about how the product’s interfaces look and function while UX, as mentioned in section 2.1, is the overall feel of the experience (Career Foundry, 2020).

2.2 Serious games

A serious game is defined as a game that is developed for a primary purpose other than enjoyment (Alvarez et al., 2010). Based on the definition, serious games are designed for education, training, scientific exploration, health care, engineering (Gaia, 2020). The word “serious” might come from the role of serious games in helping the player gain knowledge or learn new skills (Fedwa et al., 2014).

Serious games have been increasing in the games industry in recent years. Allied Market Research published a report in 2017 that estimated the worth of the serious games market worldwide was $2.7 billion in 2016 and is predicted to increase to 9.1 billion by 2023. There are three factors that play an important role in market growth: demand for customer engagement, rise in serious games which run on mobile, and improving learning results. However, the expansion of the market is hidden due to inappropriate game design and the absence of public attention to serious games. According to the report, in 2016, the area which made the highest profit from the serious games market was the Asia-Pacific. This is because that the area has a large population and a great number of people who often active on social media and on the internet. It is estimated that Latin America, Middle East, and Africa would overtake Asia-Pacific and become the areas having the highest revenue by 2023, attributed to the surge in internet usage, and changes in customer population (Figure 2). Besides, countries in the Middle East concentrate on supporting the process of digitization, soft skills development in particular (Allied Market Research, 2017.)

Figure 2. The growth of global serious games market (Allied Market Research, 2017).

2.2.1 Classification of serious games

There are many criteria that could be used to classify serious games. For example, a serious game could be classified by activity type required by the game such as running, jumping, thinking tasks. Alternatively, a game could be put in a category by its environment, for instance, two-dimensional (2D), three-dimensional (3D), virtual or mixed reality environment (Rego et al., 2010.) However, purpose-based classifications could be one of the most common ways to define types of serious games. The criteria refer to the goal of the design of the game (Djouti et al., 2011.) This section will discuss some popular areas that serious games are developed for.

The games used for education and training are the games that are designed to aid players in gaining knowledge in specific fields (Djaouti et al., 2011). There are many reasons behind applying games in the learning process. However, the most common reason could be the lessons are too long and in some cases are boring. As a result, learners have difficulty in concentration on the lessons. Thus, game-based learning could be a new way to attract learners' attention and encourage their motivation to continue the learning process (Dadheech, 2018.) Furthermore, educational games have certain

benefits. Similar to many other games, they might be addictive. While in some cases this is criticized and causes serious problems, when it comes to education, it is considered an advantage because a game that can make users feel enthusiastic to come again is a success for the designers, and from learners' point of view, the more they play, the more they learn. Moreover, serious games provide instantaneous feedback assisting users to monitor their performance. With this advantage, users can check whether they are at the right progress or falling behind. Therefore, they can adjust their activity to fit the current situation (Etherington, 2016.) Moreover, through educational games, users learn by doing. When playing a game, players have a chance to learn new rules and absorbing knowledge by interacting with the game instead of remembering them (Pears, 2016.)

Serious games are often developed for training and simulation. Similar to game-based learning, games for training and simulation have the purpose of helping users learn new knowledge, but they focus more on improving specific skills for users (Djaouti et al., 2011.) Serious games for training have the same advantages as educational games. However, the most well-known benefits of training and simulation games could be risk-free and cost-effective. Thanks to the development of virtual reality (VR), augmented reality (AR), the trainees can experience a virtual training environment that looks real. This allows trainees to practice and keep them away from unsafe situations. Because the simulators can simulate something look real, it is thus very practical to simulate the materials in virtual reality. This prevents causing any damage to the real materials (Gamelearn.) The Fire Safety Simulation, for example, is a VR gamified application created by Turku Game Lab. The application aimed to use the virtual reality environment to instruct how to correctly and safely respond to an electrical fire (Ranta, 2020.) Using the simulation might help unexperienced trainees to avoid any lethal threat that they may face in a real-life fire.

It is believed that doing regular exercise benefits human health. The exercise types could be physical, mental, or physiological (Fedwa et al., 2014.) Wellness is another popular purpose of the development of a serious game. These games are often called fitness games or exergaming and are also a kind of exercise. Exergaming has some advantages. Firstly, exergaming increases user motivation and the level of engagement with physical exercise. Motivation could be the first and the most important factor that causes someone to do exercise. If someone does not want to play sport, then it is likely that he or she will not do it. Therefore, if the level of motivation of users is increased, they will be more active and it will benefit their health. Moreover, because a fitness game

is also a game, hence it contains entertaining elements. Consequently, users find it more enjoyable. Finally, exergaming allows for personalization. This means that the game may have different instructions and targets to suit the needs of a particular user (Benzing et al, 2018.)

2.3 Methodology

This section explains all methods applied in this study. These methods are used because they are very common techniques to measure the user experience. They will directly answer the research question and define the objectives.

Questionnaires are one of a variety of methods of gathering information from people by asking indirect or direct questions. Using questionnaires brings many advantages for the research. First of all, it is one of the most low-cost ways of collecting quantitative data (Gillham, 2007.) Secondly, questionnaires are also practical. This means that the surveyor designs the questions that are asked and the range of answers that can be given. Therefore, it could help the researcher to examine his objectives (Gillham, 2007.) Furthermore, questionnaires are a useful method regarding analysis and visualization. Finally, the researcher can collect the results quickly and put the results on a wide range of charts and tables for analysis (Survey Anyplace, 2020.) Questionnaires were used as a research method in this thesis because they are inexpensive, give quick results, and the results are easy to visualize. The study estimated that over 100 people would participate in the research. Therefore, questionnaires are the best option in this case to collect information.

A case study is a research method that examines very closely a specific case or cases in a real-world situation such as an application, organization, industries (Bromley, 1986). A researcher can gather rich data by using case studies when investigating a topic. Because in-depth information is collected, case studies can eliminate the human factors such as human thinking and behavior that could not be reached in other methods (McLeod, 2019). The case study was chosen to be a method in this study because it provides a detailed understanding of the relationship between speech recognition and user experience in a serious game.

3 SPEECH RECOGNITION

3.1 Artificial intelligence and natural language processing

3.1.1 What is artificial intelligence?

The term artificial intelligence (AI) was invented by John McCarthy in 1956. He called it “the science and engineering of making intelligent machines” (Dheeraj, 2013). In general, artificial intelligence refers to the act of developing a computer, robot, or other machines which might have the ability to think like humans (Bernard, 2020). In other words, artificial intelligence helps these machines mimic the human mind such as in learning, making decisions, solving problems (IBM Cloud Education A, 2020).

There are two other common terms that are related to AI: machine learning and deep learning. Machine learning is a part of AI. Machine learning refers to computer programs that can learn by themselves without being helped by humans as inputting data increased. Deep learning belongs to machine learning. Deep learning applications teach themselves by digesting large amounts of data that is not structured such as text, images, or videos (Frankenfield, 2021.)

In the past, AI was considered science fiction, however, nowadays AI has become a part of modern life. The growth of AI development is the result of the development of computer systems that can handle data more efficiently, more accurately, and the expansion of big data. As a consequence, AI gives driving directions when being asked, suggests items customers should buy when going to markets. More importantly, the AI could make a detailed analysis to assist skilled professionals’ jobs (IBM Cloud Education A, 2020.)

3.1.2 Natural language processing

Natural language processing (NLP) is a combination of linguistics, computer science, and AI with the aim of creating the ability for computers to understand words in the form of text and spoken in the same way human beings do. With NLP, text can be translated between languages, software can react to speech commands and handle large amounts of text in real-time. The tasks of natural language processing include speech recognition, part of speech tagging, word sense disambiguation, sentiment analysis, et cetera (IBM Cloud Education B, 2020.)

Natural language understanding (NLU) is a subfield of natural language processing that deals with the task of discovering the sense of a sentence. Naturally, humans communicate with each other without caring about grammar or choice of words. However, it is very different for a computer program to do it. To do that, it requires a range of processes such as syntactic analysis, semantic analysis to know the intended meaning of a sentence (ExertAi, 2020.) It might be confusing between NLP and NLU for some people, however, the two terms are distinct from each other. NLP is a complete process of transforming the unstructured language data into structured data, then this data will be used to help machines understand and communicate with humans. While NLU only concentrates on finding the meaning of a sentence by analyzing its grammar and context (Kavlakoglu, 2020.)

3.1.3 Virtual assistant

An intelligent virtual assistant or intelligent personal assistant is a program that can recognize and understand human languages and uses its ability to complete tasks required by its users. The tasks could range from checking the weather and calendar to reading and writing text messages, making phone calls, and even control electronic devices in users' home (Hoy, 2018). The most famous assistant services include Alexa from Amazon, Siri from Apple, Cortana from Microsoft and Google has Google Assistant (Bridget, 2017). 2017 is the year seeing the expansion of using virtual assistants because of vast new products in the market and the importance of hands-free interaction. The Siri and Google Assistant are mainly used on smartphones while Cortana is installed on personal computer (PC) running with Window operating system (OS), and Amazon's Alexa is used in many smart speakers (Kline, 2017).

The development of speech recognition has a long history since the 1950s, however, applying speech recognition to assist people has just begun in recent years. It started in 2005 with the development of Siri by United State (US) Defense Advanced Research Projects Agency (McKee, 2017). Siri was then sold to Apple in 2010 and installed on iPhone 4S as a virtual assistant in 2011 (Murph, 2011). In 2016 Google Assistant was released as it could make two-way conversation (Lynley, 2016). A year later, Amazon introduced Alexa to build voice-command interfaces (Arnold, 2017).

People often misunderstand that a chatbot and a virtual assistant are the same things. However, these AI applications have some differences. A chatbot is a text-based or auditory computer program that is used to automatically communicate with humans,

instead of human with humans (Techtarget, 2017). A chatbot is usually used in dialog systems thus it is a tool for organizations to improve their customer services while the usage of a virtual assistant is for personal purposes only. Moreover, virtual assistants can understand users' moods and emotions. Although answers from chatbots are usually correct, chatbots do not remember the context of the conversations. As a result, chatbots do not react accordingly. Nonetheless, virtual assistants keep analyzing, learning after the conversations. Therefore, the talk with a virtual assistant is much more natural (Joshi, 2018).

3.2 Overview on the speech recognition

Speech is the main way of information exchange between humans. For reasons ranging from interest in technology to the desire to interacting with the machine by speech commands, the study in speech recognition has been an attractive field for a long time (Sadaoki, 2005.)

Speech recognition or speech-to-text (STT) or automatic speech recognition (ASR), as it is called, is a technology that gives a computer program the ability to transcribe humans speech into readable text (Pathak, 2010). The performance of ASR relies on algorithms combining acoustic and language modeling. Acoustic modeling builds a connection between language construct and audio signals (Bhatt et al., 2020). Language modeling a probability distribution over word sequences to detect the differences between words and phrases that sound almost the same. To make speech recognition more accurate, Markov models – a kind of a stochastic model – are applied to detect temporal patterns. Natural language processing could be used in this process to make speech recognition work faster. AI and machine learning play an important role in improving speech recognition accuracy. Because the more machine learning is used in software, the more it learns (Kiwak, 2020.)

Contrary to speech recognition, speech synthesis or text-to-speech (TTS) is a technology that receives written text as input, then outputs a speech (Allen, 1987). The synthesis process requires a large database of recorded spoken words and sentences which later is used to generate sound that is similar to a human speaking (Rubin, 1981). A speech synthesis system contains two segments: front-end and back-end. The front-end part takes raw text such as words, symbols, numbers, abbreviations and transforms them into written-out form. Phonetic transcriptions then are given to each word and the text is

divided and marked as phrases, clauses, or sentences. The output of the front-end process is symbolic linguistic representation. In the back-end stage, often called synthesizer, the symbolic linguistic representation is transformed into sound (Santen, 1994.) The level of naturalness and intelligibility is considered the key criteria in evaluating how good a synthesis system is (Taylor, 2009).

3.2.1 Brief history of speech recognition

The initial effort to design speech recognition systems was created in the mid 20th century when numerous scientists attempted to utilize the basic concept of acoustic phonetics. Since the research of programmable computing devices were in early stages of development, spectral resonances were used in most of the speech recognition study (Sadaoki, 2005.)

Speech-to-text investigations were concentrated on isolated words during the 1970s (Itakura, 1975). In the 1980s, the issue of isolated words has been solved and the system could detect a string of words. During the next decade, various changes had been made in the area of pattern recognition. The main goal in this period was to reduce the error of recognition and to give the best fitting of distribution function to given data (Juang et al, 2000).

At the beginning 21st century, artificial learning was applied in speech recognition research. In 2005 some innovations were made in automatic speech recognition as long speech as continuous speech recognition and performance improvements on large scale speech database (Giuseppe, 2005). In 2017, there was a milestone for speech recognition as researchers from Microsoft successfully transcribing a phone call conversation on a popularly benchmarked Switchboards task. Deep learning models were also applied to improve the accuracy of recognition. As a result, the rate of error was described to be equal to 4 speech transcribing experts working together on the same benchmark (Huang, 2017.)

3.2.2 Speech recognition developed by some companies

In terms of commercials, the most used speech recognition services are from well-known companies. Google has “Google Web Speech application programming interface”, IBM developed “Watson speech-to-text service” and Microsoft created “Azure Speech Services” (Aguilar-Chacon et al., 2020.)

IBM is one of the most well-known and oldest companies in computer science in the United State. In terms of researching speech recognition, it was the pioneer in the field as it was the first company that successfully developed the speech recognition system in 1962. In 2011, the Watson machine was released. The machine can understand natural language and communicate with humans (IBM Team.) Currently, everyone can use IBM's speech recognition service through Watson cloud services. The speech-to-text service supports 11 languages in real-time or from audio files with different formats such as Wav, Mp3, Mpeg files. Regarding programming, IBM provides application programming interface (API) that can be implemented with C#, Java, Python, et cetera (IBM Cloud a, b, c.)

Google is another popular company in the technology field. Although it was founded just over 20 years ago, it has become one of the biggest companies in the world nowadays (Mahesh, 2021). In 2012 Google launched Google Voice Search that used speech recognition to help users search for information. Now speech recognition is separated from Voice Search and available on Google Cloud Services (Bertrand, 2011.) The service fully or partially supports 120 languages. The service can convert speech to text in real-time or from prerecorded audio (Lawton, 2019.) Google also offers APIs in programming languages such as C#, C++, Java, Go (Google Speech-to-Text c).

Microsoft is also one of the leading companies in developing speech recognition technology. In 2007, for the first time, speech recognition was implemented in a Windows operating system – Windows Vista (Shinder, 2007). Now users can access the speech-to-text service through Azure cloud services, the service is available in 31 languages. Similar to other providers, the service can be used in real-time or with prerecorded audio. With APIs provided by Microsoft, developers can program in C#, C++, Java, et cetera (Mircosoft Docs a, b, c.)

Table 1. Summary of speech recognition services (IBM Cloud a, b, c; Google Speech-to-Text a, b, c; Mircosoft Docs a, b, c).

	Goolge	IBM	Microsoft
Programming language, engine support	C#, C++, Java, Python, Go, Node.js, PHP, Ruby	C#, Java, Node.js, Python, Go, Ruby, Swift, Curl, Unity	C#, C++, Java, Python, Go, Node.js, JavaScript, Swift,
Audio format support	Mp3, Flac, Linear16, Mulaw, Amr, Amr_wb, Ogg/Opus	Wav, Mp3, Mpeg, Ogg, Webm, Alaw, Flac, Mulaw	Mp3, Opus/Ogg, Flac, Alaw, Mulaw
Language support	120	11	31

In a research conducted by Aguilar-Chacon and Segura-Torres, the researchers examined speech recognition services from IBM, Google, and Microsoft to discover which service having the best result in conversation time and precision. The experiment was done on Raspberry Pi 3 with the Spanish language. In the experiment, 10 recorded audio files were deployed on the three services and the experiment was repeated several times. The results illustrated that Google's system was the fastest and IBM's was the slowest. As shown in table 2, Google speech to text's average conversation time was 6101,579 millisecond, Microsoft's was 9205,645 millisecond, and IBM's was 11744,616 millisecond (Aguilar-Chacon et al., 2020.)

Table 2. General average of time, with respect to STT systems (Aguilar-Chacon et al., 2020).

STT systems	Average conversion time (mS)
Google	6101,579
Microsoft	9205,645
IBM	11744,616
Grand Total	9017,279822

The research also showed that the results from Google's service were 97,116% similar to the context of audio inputs. While Microsoft's and IBM's results were 90,825% and 90,337% respectively as displayed in table 3 Aguilar-Chacon et al., 2020.)

Table 3. General average of precision in percentage, with respect to STT systems (Aguilar-Chacon et al., 2020).

STT Systems	Similarity percentage Average
Google	97,116%
Microsoft	90,825%
IBM	90,337%
Grand Total	92,759%

In another research done by Filippidou and Moussiades, it aimed to find the word error rate (WER) of three speech recognition services from Google, IBM, and Wit (Filippidow et al., 2020). WER is used to measure how good a speech recognition system perform (Gevirtz, 2018). The test was conducted in English with three different speakers whose mother tongue is not English. The results of the research showed that Google's service outperformed IBM's (Filippidow et al., 2020.)

Table 4. Average WER (%) measurements (Filippidou et al., 2020).

Speaker A			**Speaker B**			**Speaker C**		
IBM	**Google**	**Wit**	**IBM**	**Google**	**Wit**	**IBM**	**Google**	**Wit**
30.10	16.60	25.87	47.73	20.45	23.28	36.51	24.85	58.87

Technically, Google's speech recognition seems better than IBM's and Microsoft's as it supports vast languages and having good results in experiments. However, IBM's speech recognition was chosen in this thesis because it could help reduce the development time. IBM provides SDK for Unity with detailed instruction. Unity is a game engine that was used to develop ExerGo. Plus, the SDK has examples of how to implement Watson's services in Unity that programmers can learn and apply in their projects. On the other hand, Google and Microsoft services do not offer a similar approach to Unity as IBM does. If Google or Microsoft services had been chosen, the speech recognition feature would have been built from scratch which could raise the development time. Moreover, the author did not feel confident to develop such a complex

feature like speech recognition from scratch. Therefore, by using IBM speech recognition, it could reduce the development time.

3.3 Speech recognition applications

Since speech recognition became more and more popular in everyday life, it brings a new kind of application which is very different from other computer application. It gives users an easy way to interact with software without physical interaction. With that advantage, speech recognition is applied in many fields of technology.

The area which might benefit the most from speech recognition is handicapped and disabled people. For the ones who are deaf or difficult of hearing, they can understand a conversation by using speech recognition software which transcribes the conversation into text (MassMatch, 2010). Speech recognition is one of the alternative options for people who are unable and difficult to use their hands to access the computer system (Pathak, 2010).

Speech recognition is also applied in the healthcare system to help doctors. The ability to translate spoken words into text and reading information allows physicians to communicate with the digital health record system by purely saying some commands. Speech recognition could be used for dictation to help typists. Speech recognition researches are continuing and used parts in the healthcare system. Dragon Naturally Speaking Medical Solutions, for example, is a software developed that can translate a doctor's voice into a detailed narrative that helps cut down the time for documentation (Pathak, 2010.)

Using speech recognition applications is also noticed in the field of aviation. The speech recognition system in helicopters has been examining in a joint program conducted by the US Army Avionics Research and Development Activity and by the Royal Aerospace Establishment. The result is wonderful and promising. The Puma helicopter from France has also applied speech recognition to the system of navigation setting, control of communication radios, and control of an automated target handover system. Additionally, speech recognition is used in a flight simulator and for training such as the Microsoft Flight Simulator program and voice-driven system training in the US military (Clifford, 1990.)

Mobile devices and computers could be the most notable area for speech recognition applications. Speech recognition is used on these devices for practical things. Windows 10, the newest version of the operating system from Microsoft for PC and laptop, speech recognition is used for navigating, dictation, and other basic commands such as copy, paste, delete, et cetera (Microsoft, 2020.) On the latest iPhone generation, for example, users can make a call, message, set an alarm clock, open apps, remind something… through speech recognition. Moreover, it contributes to users' safety when the user can interact with the phone without touching it. This is very important when the user is on the move. As the user is driving alone, for instance, and he or she receives a new message on an iPhone, he or she could say "Read my new messages", then the phone reads them loudly (Pogue, 2020).

3.3.1 Speech recognition in games

This section of the thesis will introduce the development of speech recognition in the game industry.

There are many levels of how speech recognition is applied in games. In some games, speech recognition is just any human's voice, in other games, it requires more complex and accurate speech recognition to play the game. In Chicken Scream – a mobile game released in 2017, for example, a normal talking voice is the signal to move the chicken, by contrast, a sudden and loud voice makes the chicken jumping. Bot Colony is another example of a game that requires a complex system of speech recognition where players need to communicate with the game by voice commands (Kiiski, 2020.)

The first attempt to develop voice interaction games started in the 1960s when researchers used a few standardized words to control simple games for testing. As a result of much research, in 1973 Voice-Chess was released as the first game that implemented speech recognition technology. Hearsay-I, a speech recognition system, was applied in the game to know and react to speech commands such as "Bishop to Queen Three". The Hearsay-I understanding system received spoken words and made some signals based on the input (Reddy, 1973.)

The Year 1983 was considered a landmark for speech recognition games as there were some efforts to make those games commercial products. A video console game from RDI Video System named Halcyon which was predicted the most successful game on the market. The game had some features which required a speech recognition system

such as the game could be controlled by spoken language, detecting a particular player by his or her voice, and the ability to add new words. However, the attempt to bring the game to the market failed followed the crash of the North American video game in 1983. Contrary to North America, in Japan speech recognition games had some success with the born of Famicom in 1983. Famicom was a video game console system founded by Nintendo, the platform had two controllers, one of them had a microphone integrated for speech recognition use. However, speech recognition games were not popular at that time therefore the new feature was ignored by players and was removed in the next version of the console (Fraser, 2017.)

Over the next two decades, speech recognition games were developed in two different ways between Japan and North America developers. Japanese approach was to encourage the users' enthusiasm for the speech recognition experience. They concentrated on two-way, conversational speech recognition games between the player and the game characters. By contrast, North American game developers weighted to the one-way conversation as the player giving speech commands to game characters.

In parallel, karaoke, a digital game genre that had voice interaction was considered very successful at the time. The elements separated karaoke from other games having voice interaction were the stage-style microphone using as a symbol and the role of the game in the society as social linkages were strengthened (Fletcher and Light, 2011.)

Voice interaction technology was one step closer to be used in-game as interfaces because of the born of the Microsoft Kinect. This had led to one of the largest waves of games to feature voice interaction in some capacity. This could be the result of voice interaction built-in support devices in which the requirement for game developers to contribute their own software of speech recognition was removed (Fraser, 2017.)

Not long ago, many independent game developers have been trying to add voice interaction into existing games. This is particularly notable as the first wave of voice interaction games to have arisen primarily on computers and mobile phones rather than consoles, enabled by a series of developments in hardware and software availability. Comparing these recent projects to games of the past, independent developers have displayed a particularly focused engagement with the design challenges and opportunities that are inherent to voice interaction in a gaming context (Fraser, 2017.)

4 EXERGO GAME

4.1 The goal of ExerGo game

The ExerGo is a mobile health application running on Android OS that monitors some information while users running. The app purposes to motivate users to do physical exercise more. The ExerGo is developed by developed by Turku University of Applied Sciences namely in Futuristic Interactive Technologies research group as a part of Business Finland funded project called Business Ecosystems in Effective Exergaming (BEE). The BEE project was designed to make the quality of exercise games better and demonstrate their usefulness and effectiveness. The BEE project also purposes to raise the level of attractiveness and motivation effect of exercise games.

Figure 3. ExerGo running mode

The app requires the device to have internet-connected and Suunto Movesense sensor to measure the heart rate. When ExerGo first installed, there is a list of questions that ask users about users' health conditions to ensure users can take the tasks. After that, the user can choose the level of exercise from the low level to the high level. The higher the level, the longer the exercise is. For example, the beginning level requires a user to run for 20 minutes, but with an advanced level, the user needs to run for 40 minutes. There are three phases for each run: warm-up, work, and recovery. Each state requires a different speed, for instance, the warm-up state is normal running, the work state is faster and the last state slows down the speed. Plus, the total time will be divided according to phases. After each run, the statistics such as the average speed, the completed distance, the heart rate will be saved and shown to the user.

4.2 Implementing speech recognition

When using the app, it is difficult for users to check the device's screen and in some circumstances it could be dangerours. It needs another way for users to monitor the process. Therefore, FIT researchers decided to use speech recognition. As mentioned in section 3.2.2, the speech recognition system used in ExerGo is from IBM because IBM provides SDK tools for Unity engine.

4.2.1 Preparation

After users pressing the 'Start' button on the app, the app will enter the running mode as shown in figure 3. Here speech recognition will be implemented and help users to get updated information such as the time left, the current heart rate, the distance has been running…

There are two things needed to use IBM Watson SDK in Unity. Firstly, the IBM Watson SDK must be installed into the current project. The SDK can be downloaded from the GitHub repository of Watson developer cloud. Depending on the Unity version, it requires updating the build setting in Unity to .NET 4.x equivalent. Secondly, developers need to have rights to access Watson cloud services by creating an account. After creating an account, the users can register the services they need and get the credentials for each service. To access Speech-to-Text and Text-to-Speech services it needs to have API and uniform resource locator (URL) that can be found from the IBM cloud dashboard (Mikemosca, 2021.)

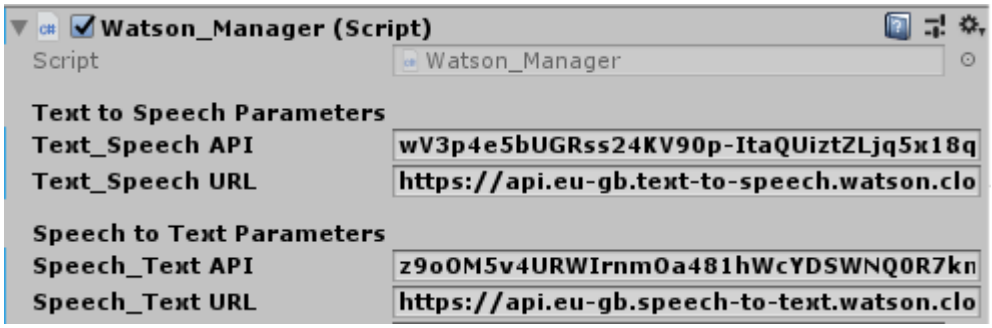

Figure 4. APIs and URLs in Unity inspector

Figure 5 is the flowchart of how the speech recognition feature was created. When running mode is started, at the same time, Speech-to-Text and Text-to-Speech services started meaning it will make a connection to IBM Watson cloud. The Speech-to-Text will continuously listen to any spoken word except when audio from Text-to-Speech is playing. If the Speech-to-Text script detects any human speech, it will transcribe and return a string to the Watson-Manager script. The Watson-Manager will compare the string with a list of defined commands. If the string matches one of the commands, then Watson-Manger will take data from the app and convert it into a string. Otherwise, Watson-Manager will return the “I don’t understand” string. The Text-to-Speech script will take string input from Watson-Manager, convert the string into audio and play it. After the audio playing, the process returns to the Speech-to-Text script.

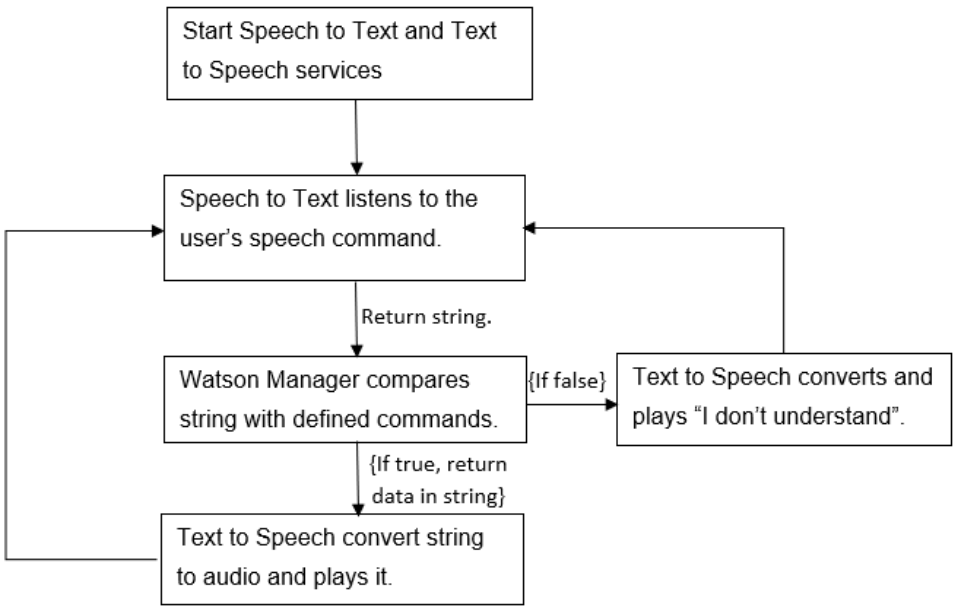

Figure 5. Flowchart of speech recognition feature

4.2.2 Speech-to-Text

To use Speech-to-Text service, first it needs to connect to IBM cloud service. The Speech-toText script takes API and URL from Watson-Manager script, then makes a request to IBM cloud. The IBM cloud will check the request and send a tocken back. If API and URL are correct, the tocken is the permission to access the Speech-to-Text service, otherwise, the tocken will be empty.

```
private IEnumerator CreateService() {
        IamAuthenticator authenticator = new IamAuthenticator(apikey: _iamApi
key);
        //  Wait for token
        while (!authenticator.CanAuthenticate()) {
            yield return null;
        }
        _service = new SpeechToTextService(authenticator);
}
```

Once the connection is made, the service can be used. It will continuously listen to any sound from the microphone. If it recognizes the sound is human speech, then the sound will be recorded. The recording method is designed to record only when no audio clip is playing to avoid a possible loop. The recording method will automatically stop when there is silent or too long speaking. The recording then will be sent to IBM cloud to be analyzed and transcribed.

When the result ready, it can be accessed by using a callback method. The method will return a string and directly send to Watson-Manager script.

4.2.3 Text-to-Speech

After receiving the string from the Speech-to-Text script, the Watson-Manager will compare the string with a list of defined commands as shown in the snippet bellow and return the result of true or false.

```
listCommands = new List<string>() {
                                    "time", "state",
                                    "heart rate", "cycle",
                                    "distance", "estimated distance"
};
```

If it is the true result, another method will be called to extract information according to the command, then convert data to a string and send it to Text-to-Speech script. If it is the result of false, a string of “I don’t understand” will be sent to the Text-to-Speech script.

Similar to Speech-to-Text, Text-to-Speech needs to be connected to IBM cloud service first and an accepted tocken to use the service. However, the API and URL of Text-to-Speech are completely different from Speech-to-Text’s API and URL.

The text received from the Watson-Manager script will be synthesized along with callback and information needed in a method in the Text-to-Speech script. After synthesizing, a clip is sent back to the Text-to-Speech script and will be played as shown in snippet bellow.

```
private IEnumerator Synthesize(string command) {
        byte[] synthesizeResponse = null;
        AudioClip clip = null;
        service.Synthesize(
            callback: (DetailedResponse<byte[]> response, IBMError error) =>
{
                synthesizeResponse = response.Result;
                clip = WaveFile.ParseWAV("clip", synthesizeResponse);
                PlayClip(clip);
            },
            text: command,
            voice: voiceType,
            accept: synthesizeMimeType
        );

        while (synthesizeResponse == null)
            yield return null;
}
```

5 RESULTS AND FINDINGS

The following section will show the results of questionnaires using in the research. Furthermore, the results will be discussed and analylized.

5.1 Results

Questionnaire 1

The first survey aimed to discover what people think of speech recognition. The questionnaire is a web-based Google Forms questionnaire. The questionnaire was posted on several Facebook public groups. Overall, there were 98 people answering the questionnaire. Questionnaire 1 and its results are shown in the appendix.

Questionnaire 2

This questionnaire was used for testers of the ExerGo application. There were 22 people who tested the app. These testers were non-native English speakers and also participated in the first questionnaire. Due to the pandemic situation, all testers wanted to do the testing alone. Therefore, the app was sent to testers and gave them the instruction. Testers installed the app on their devices for testing. Because no one had the sensor to take the heart rate, thus, the heart rate was fixed for testing purposes. Every tester has 5 minutes to test the app.

When the player is ready to run, the player presses the ‘start' button on the app and the app will enter the running mode. In the running mode, the progress of the player is updated simultaneously and the data is displayed on the phone screen or can be transcribed into audio. In this research, the testers were required to walk or run and have to use speech recognition to check their progress by speaking a specific speech command without looking at phones’ screens. The app would respond by audio according to which speech command was spoken. An online questionnaire was sent to players once they finish the testing. The second questionnaire and results can be found in the appendix.

5.2 Findings

The following section will analyze the results of the research. Furthermore, it will find the answer for the main research question by answering four sub-questions introduced in chapter 1.

The results of the first questionnaire illustrate that participants in the research are young because people who are from 15 to 35 years old are 88% of all participants. The student group is the largest group with over 58% of all respondents. There is little difference in participants by gender as male participants are 49.5%, female participants are 43.44%. In addition, the research demonstrates that technology and business are the most popular fields which nearly 70 participants studied or are studying. The survey discovered that the habit of playing games is popular among people. It is because in the survey people who spend 5 to 15 hours and 15 to 25 hours per week about 55% of all participants. Besides multiplayer online battle games, role-playing games, action-adventure games are the most favored types of games in this survey. Those types of games together have attracted 60 of 98 participants playing them. The thesis did not compare the popularity of serious games with other types of games. However, the study discovered that 76 participants have played at least one serious game that is equivalent to 77% of all participants. Moreover, many serious games which participants have played are effective. In fact, around 50 out of 98 participants thought the serious games have helped them to gain new knowledge or learn new skills.

Furthermore, the results illustrate that using the speech recognition system is not a usual thing when the number of people who rarely or never use it is nearly 50% of all participants. Nonetheless, a large majority of participants are aware of speech recognition as about 85% of participants have used it at least once. In addition, speech recognition technology seems to be effective. People in the survey considered using speech recognition because of productivity other than entertainment or curiosity. Moreover, the study finds that the speech recognition system is helpful among many participants as the number of people who thought so is bigger than the people who had different opinions.

The discussion in this section has answerd the first sub-question of the research: what do people think of speech recognition? Overall, participants had positive opinions on

speech recognition, and it seems that they have been using it because of its effectiveness.

5.2.1 Advantages of speech recognition in serious games

In this part, the consequences of question 3, 6, and 7 in the second questionnaire will be examined to find some advantages of speech recognition in serious games were discovered.

Firstly, using the speech recognition system helps players to concentrate. After testing the ExerGo game, 77% of players responded that they do not pay attention much to communicate with the game while still focusing on completing tasks in the game. This is because the ExerGo is a physical game that requires the player to run, if players check the game's information by looking at the phone's screen, players could be distracted from moving on the street. Thus, by using speech recognition, players can give speech commands and audio with information will inform the players.

Secondly, the voice-user-interface from the speech recognition received good feedback from players. In the survey, major responses to question 6 of the second survey answered that they enjoyed hand-free interaction when playing the game. Because users just speak speech commands to get information, it could give users more time and concentration to do other tasks rather than looking and touching at phones' screens. Thus voice-user-interface could play an important role in helping the player to immerse in serious games.

Moreover, results from question 7 demonstrate that many players did not bored with speech recognition. This could be a sign that speech recognition could be implemented more on serious games.

In general, there are two main benefits of speech recognition in serious game: helping users to concentrate and giving more time for users to do other tasks. This conclusion has provided the answer for the research sub-question: what are the benefits of implementing speech recognition to such project?

5.2.2 Disadvantages of speech recognition in serious games

On the other hand, the second questionnaire found some drawbacks of speech recognition on the ExerGo game. This section will answer the research question: what are the problems users have when they test the game?

First of all, the speech recognition feature responded slowly to the players' speech commands. The outcomes of question 4 in the second questionnaire display that the number of participants who believed the speech recognition feature being laggy is bigger than the participants who opposed it by 15%. This means players did not satisfy when using speech recognition to play the game. In the process of developing a game or an application, developers always find a way to minimize the latency of the game or the application because no one wants to use a slow application in real life. The problem of delay could be caused by long speech commands because long speech commands might take a longer time to process. Thus, implement short speech commands could improve the.

The next factor which made a negative effect on users' experience while playing the Exergo game was the inability of the speech recognition feature to recognize spoken commands from some users. According to results from question 5 of the second questionnaire, the number of participants who thought the game did not understand their spoken words is bigger than the number of people who thought the game knew what they spoke. This is a crucial factor because if the speech recognition feature does not know what users speak, then the programmed functions relying on the speech recognition system do not work. As a result, the speech recognition feature might be useless and a bad user experience will be inevitable. There might be many reasons behind the issue. A small voice, for example, could cause the problem because the speech recognition could not detect a speech command or could not distinguish between a human voice and other sounds. Another reason may be testers are not English native speakers, thus, the commands might not be pronounced correctly leading to speech recognition not understanding what testers spoke.

In conclusion, the answer of the question: "what are the problems users have when they test the game?" are the latency of speech recognition and its ability to understand speech command.

5.2.3 Factors affecting user experience in serious games

Question 1 and question 2 in the second questionnaire were purposed to find answers for the research sub-question: what factors could improve the user experience in serious games? From the results of the two questions, the usability of speech recognition and its ability to make human-like voice are the factors could have a good impact on user experience.

The first factor is the usability of the speech recognition feature. The results of the first question revealed that the speech recognition system was easy to use. Usability is an important element enhancing the user experience. Thus, the high rate of positive answers to question 1 demonstrates users' satisfaction while playing the Exergo game.

Additionally, interacting with the game via a speech recognition system made players feel human-like conversation. This can be proved by the results from question 2 of the second questionnaire. The results illustrate a large percentage of players feeling the conversations were natural as talking to a human. For that reason, using speech recognition could help players feel comfortable, and less stressed while interacting with the game.

6 CONCLUSION

The thesis aimed to investigate how participants thought of speech recognition and explore the factors that affect user satisfaction when users playing a serious game with speech recognition. Moreover, the research discovered the benefits and drawbacks of a speech recognition system in a serious game. All the goals of this thesis were achieved.

The ExerGo was chosen as a case study and IBM speech recognition was implemented on ExerGo. Two questionnaires were used in the research. The first questionnaire aimed to examine participants' opinions on speech recognition. The second questionnaire was used after participants tested the ExerGo game. The purpose of the second questionnaire was to explore factors of speech recognition that influenced user experience, advantages, and disadvantages of speech recognition in the ExerGo game.

The results from the research suggested that participants have a positive view of speech recognition. Testers seemed like the speech recognition because it helped them to concentrate and hands-free interaction giving them a convenient feeling. The research also showed that the usability of speech recognition and the human-like conversation could be the factors that affect user experience. Furthermore, the study discovered that the inability of speech recognition to work faster and more accurate might have a negative influence on user experience.

However, there was a limitation in the research. Because of the pandemic, no tester had the Suunto Movesense sensor to measure heart rate, therefore, the heart rate on the ExerGo app was fixed at a specific number. This could have reduced the user experience and the result of the research could have been different.

Further research could focus on investigating the solutions for the problem of delay of speech recognition found in the research. Another direction could be examining why sometimes speech recognition does not understand speech commands and finding a way to rectify this issue.

REFERENCES

Aguilar-Chacon, J.E., Segura-Torres, D.A. (2020). Evaluation methodology for Speech To Text Services similarity and speed characteristics focused on small size computers

Allen, J., Hunnicutt, S. M., Klatt, D. H. (1987). From Text to Speech: The MI Talk system

Alvarez, J., Djaouti, D., Jessel, J.-P., O. Rampnoux. (2010). Origins of serious games

Benzing, V., Schmidt, M. (2018). Exergaming for children and adolescents: Strengths, weaknesses, opportonities and threats

Bhatt, S., Jain, A. (2020). Acoustic Modeling in Speech Recognition: A Systematic Review

Bromley, D.B. (1986). The Case-Study Method in Psychology and Related Disciplines

Clifford J. W. (1990). Opportunities for Advanced Speech Processing in Military Computer-Based Systems

Dheeraj, S., Ajitanshu, V., Gyanendra, S. (2013). An overview of artificaial intelligence.

Djaouti, D., Alvarez, J., Jessel. (2011). Clssifying Serious Games: the G/P/S model

Fedwa, L., Mohamad, E., Abdulmotaled, E. S. (2014). An overview of serious games

Filippidou, F., Moussiades, L. (2020). A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems

Fraser, A. (2017). A History of Voice Interaction in Games

Gillham, B. (2007). Developing a Questionnaire, 2nd edition

Giuseppe, R. (2005). Active Learning: Theory and Applications to Automatic Speech Recognition

Hoy, M.B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants

International Organization for Standardization. (2019). Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en

Itakura, F. (1975). Minimum Prediction Residula Applied to Speech Recognition

Juang, B.H., Furui, S.. (2000). Automatic speech recognition and understanding: A first step toward natural human machine communication

Kiiski, T. (2020). Voice Games: The History of Voice InteractioninDigitalGames

Luchs, G., Griffin, A. (2016). Design Thinking: New product development. John Wiley & Sons. New Jer-sey

McKee, Heidi. (2017). Professional Communication and Network Interaction: A Rhetorical and Ethical Approach

Pathak, P. (2010). Speech recognition technology: Application & future

Ranta, J. (2020). A comparative study of hand tracking in a VR environment

Rego, P.A, Moreira P.M, Reis L.P. (2010). Serious Games for Rehabilitation: A survey and a classification towards a taxonomy

Rubin, P., Baer, T., Mermelstein, P. (1981). An articulatory synthesizer for perceptual research

Sadaoki F. (2005). 50 years of Progress in speech and Speaker Recognition Research

Santen, V.J. (1994). Assignment of segmental duration in text-to-speech synthesis

Taylor, P. (2009). Text-to-speech synthesis

Online sources:

Allied Market Research. (2017). Serious games market outlook: 2023. [online]. (Last updated 12 Dec 2020). Available at: https://www.alliedmarketresearch.com/serious-games-market

Arnold, S.E. (2017). Bradley Metrock and the Alexa Conference: Alexa As a Game Changer for Search and Publishing. [online]. (Last updated 7 Apr 2021). Available at: http://arnoldit.com/wordpress/2017/02/02/bradley-metrock-and-the-alexa-conference-alexa-as-a-game-changer-for-search-and-publishing/

Bernard, M. (2020). What is AI. [online]. (Last updated 7 Apr 2021). Available at: https://www.bernardmarr.com/default.asp?contentID=963

Bertrand, D. (2011). Introducing voice actions for android in the UK, France, Italy, Germany and Spain. [online]. (Last updated 10 Apr 2021). Available at: http://googlemobile.blogspot.com/2011/09/introducing-voice-actions-for-android.html

Bridget, B. (2017). Virtual assistant. [online]. (Last updated 7 Apr 2021). Available at: https://searchcustomerexperience.techtarget.com/definition/virtual-assistant-AI-assistant

Bloomberg. (2019). More than 150 Voice Technology Companies to Showcase Latest Innovations in Artificial Intelligence and Voice AI Soluti. [online]. (Last updated 11 Mar 2021). Available at: https://www.bloomberg.com/press-releases/2019-07-19/more-than-150-voice-technology-companies-to-showcase-latest-innovations-in-artificial-intelligence-and-voice-ai-soluti

Cameron S. (2020). The importance of user experience. [online]. (Last updated 25 Nov 2020). Available at: https://www.vincit.com/blog/the-importance-of-user-experience

Career Foundry. (2020). The Difference Between UX And UI Design. [online]. (Last updated 20 Nov 2020). Available at: https://careerfoundry.com/en/blog/ux-design/the-difference-between-ux-and-ui-design-a-laymans-guide/

Dadheech, A. (2018). The importance of Game Based Learning in Modern Education. [online]. (Last updated 24 Apr 2021). Available at: https://theknowledgereview.com/importance-game-based-learning-modern-education/

Etherington, C. (2016). Top 8 benefits of serious games. [online]. (Last updated 24 Apr 2021). Available at: https://www.eleapsoftware.com/top-8-benefits-of-serious-games/

ExpertAi. (2020). Natural language understanding: whai is it and how is it different from NLP?. [online]. (Last update 25 Apr 2021). Available at: https://www.expert.ai/blog/natural-language-understanding-different-nlp/

Frankenfield, J. (2021). Artificial intelligence. [online]. (Last updated 7 Apr 2021). Available at: https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

Gaia. (2020). Serious games. [online]. (Last updated 11 Mar 2021). Available at: https://cs.gmu.edu/~gaia/SeriousGames/index.html

Gamelearn. Eight examples that explain all you need to know about serious games and game-based learning. [online]. (Last updated 24 Apr 2021). Available at: https://www.game-learn.com/all-you-need-to-know-serious-games-game-based-learning-examples/

Gevirtz, M. (2018). What is Word Error Rate (WER)?. [online]. (Last updated 25 Apr 2021). Available at: https://deepgram.com/blog/what-is-word-error-rate/

Google Speech-to-Text a. Speech-to_text documentation. [online]. (Last updated 25 Apr 2021). Available at: https://cloud.google.com/speech-to-text/docs/languages

Google Speech-to-Text b. Speech-to_text documentation. [online]. (Last updated 25 Apr 2021). Available at: https://cloud.google.com/speech-to-text/docs/encoding

Google Speech-to-Text c. Speech-to_text documentation. [online]. (Last updated 25 Apr 2021). Available at: https://cloud.google.com/speech-to-text/docs/client-libraries

Huang X. (2017). Microsoft researchers achieve new conversational speech recognition milestone. [online]. (Last updated 1 Dec 2020). Available at: https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/

IBM Cloud a. Speech to text. [online]. (Last updated 10 Apr 2021). Available at: https://cloud.ibm.com/docs/speech-to-text?topic=speech-to-text-models

IBM Cloud b. Speech to text. [online]. (Last updated 25 Apr 2021). Available at: https://cloud.ibm.com/apidocs/speech-to-text?code=unity

IBM Cloud c. Speech to text. [online]. (Last updated 25 Apr 2021). Available at: https://cloud.ibm.com/docs/speech-to-text?topic=speech-to-text-audio-formats

IBM Team. Pioneering speech recognition. [online]. (Last updated 11 Mar 2021). Available at: https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/

IBM Cloud Education A. (2020). Artificial Intelligence. [online]. (Last updated 7 Apr 2021). Available at https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

IBM Cloud Education B. (2020). Natural Language Processing. [online]. (Last updated 25 Apr 2021). Available at: https://www.ibm.com/cloud/learn/natural-language-processing

Interaction Design Foundation (2020). The 7 Factors that Influence User Experience. [online]. (Last updated 11 Nov 2020). Available at: https://www.interaction-design.org/literature/article/the-7-factors-that-influence-user-experience

Joshi, N. (2018). Yes, Chatbots and virtual assistants are different!. [online]. (Last updated 8 Apr 2021). Available at: https://www.forbes.com/sites/cognitiveworld/2018/12/23/yes-chatbots-and-virtual-assistants-are-different/?sh=2ce1939a6d7d

Kavalkoglu, E. (2020). NLP vs. NLU vs. NLG: the differences between three natural language processing concepts. [online]. (Last updated 25 Apr 2021). Available at: https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

Kiwak, K. (2020). Speech recognition. [online]. (Last updated 25 Apr 2021). Available at: https://searchcustomerexperience.techtarget.com/definition/speech-recognition

Kline, D.B. (2017). Alexa, how big is Amazon's echo?. [online]. (Last updated 7 Apr 2021). Available at: https://www.fool.com/investing/2017/01/30/alexa-how-big-is-amazons-echo.aspx

Lawton, G. (2019). A comparison of 6 speech-to-text services. [online]. (Last updated 25 Apr 2021). Available at: https://searchcloudcomputing.techtarget.com/tip/Evaluate-speech-to-text-services-from-AWS-Microsoft-and-Google

Lynley, M. (2016). Google unveils Google Assistant, a virtual assistant that's a big upgrade to Google Now. [online]. (Last update 7 Apr 2021). Available at: https://techcrunch.com/2016/05/18/google-unveils-google-assistant-a-big-upgrade-to-google-now/

Mahesh, M. (2021). Over 271 Google products & services you probably don't know. [online]. (Last updated 10 Apr 2021). Available at: https://www.matrics360.com/google-products-and-services/

MassMatch. (2010). Overcoming Communication Barriers in the Classroom.[online]. (Last updated 2 Dec 2020). Available at: https://www.massmatch.org/aboutus/listserv/2010/2010-03-31.html

McLeod, S. (2019). Case Study Method. [online]. (Last updated 30 Mar 2021). Available at: https://www.simplypsychology.org/case-study.html

Meticulous Blog. (2020). Top 10 companies in speech and voice recognition market. [online]. (Last updated 11 Mar 2021). Available at: https://meticulousblog.org/top-10-companies-in-speech-and-voice-recognition-market/

Microsoft. (2020). Use dictation to talk instead of type on your PC. [online]. (Last updated 15 Dec 2020). Available at: https://support.microsoft.com/en-us/windows/use-dictation-to-talk-instead-of-type-on-your-pc-fec94565-c4bd-329d-e59a-af033fa5689f

Microsoft Docs a. About the Speech SDK. [online]. (Last updated 25 Apr 2021). Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/language-support

Microsoft Docs b. About the Speech SDK. [online]. (Last updated 25 Apr 2021). Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-use-codec-compressed-audio-input-streams?tabs=debian&pivots=programming-language-csharp

Microsoft Docs c. About the Speech SDK. [online]. (Last updated 25 Apr 2021). Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-sdk?tabs=windows%2Cubuntu%2Cios-xcode%2Cmac-xcode%2Candroid-studio

Mikemosca. (2021). IBM Watson SDK for Unity. [online]. (Last updated 26 May 2021). Available at: https://github.com/watson-developer-cloud/unity-sdk

Murph, D. (2011). iPhone 4S hands-on!. [online]. (Last updated 7 Apr 2021). Available at: https://www.engadget.com/2011-10-04-iphone-4s-hands-on.html?guce_referrer=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnLw&guce_referrer_sig=AQAAAHgIYoRqiUEm_6F2s3mraknNPEWniR-UEDEOmWsD8EgopVlp_2jm1n_kaVJuV8HOMUFu5uJcTYnmCJcLL9Hjp1JJWD5IOZh-0_na_bVISHFfeEEWmQwM5PJfaBWHxJlZeuNavhGpcAR3LVvDSEy9FW9qR8nvBXHpEjTincUBFcXG

Nielsen, J. (2012). Usability 101: Introduction to Usability. [online]. (Last updated 23 Apr 2021). Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Pears, A. (2016). Why serious games? 6 Key benefits. [online]. (Last updated 24 Apr 2021). Available at: https://www.theaccessgroup.com/blog/dlc-why-serious-games-6-key-benefits/

Pogue, D. (2020). The voice-off: Android vs Siri. [online]. (Last updated 11 Mar 2021). Available at: https://www.nytimes.com/2013/08/22/technology/personaltech/android-vs-siri-the-voice-recognition-sequel.html

Survey Anyplace (2020). 10 Advantages and Disadvantages of Questionnaires. [online]. (Last updated 22 Nov 2020). Available at: https://surveyanyplace.com/questionnaire-pros-and-cons/

Shinder, D. (2007). Speech recognition in Windows Vista. [online]. (Last updated 25 Apr 2021). Available at: https://www.techrepublic.com/article/speech-recognition-in-windows-vista/

Web Designer Depot. (2009). Interview with web usability guru,
 Jakob Nielsen.[online]. (Last updated 23 Apr 2021). Available at: https://www.webdesignerdepot.com/2009/09/interview-with-web-usability-guru-jakob-nielsen/

Appendix

Questionnaire 1 results:

1. What is your age?				
Age:	15 – 25	25 – 35	35 – 45	Over 45
Answers:	47	40	11	0
Total responses:	98			

2. What is your gender?			
Gender:	Female	Male	Prefer not to say
Answers:	43	49	6
Total responses:	98		

3. Choose the option which best describes you: You are				
	a student	an employed person	an unemployed person	a freelancer
Answers:	58	15	14	11
Total responses:	98			

4. Which field did you study or are you studying?				
	Technology	Business	Natural Sciences	Philosophy
Answers:	35	34	13	8
	Arts	History	Medicine	Other
Answers:	4	3	1	0
Total responses:	98			

5. How often do you play games?				
Hours/week	1 – 5	5 – 15	15 – 25	Over 25
Answers:	40	32	22	4
Total responses:	98			

6. Which type of games do you play most?				
	Action-adventure	Escape	Role-playing	Fighting
Answers:	19	9	20	15
	Multiplayers	Shooting	Sports	Casual
Answers:	21	6	5	2
Total responses:	97			

7. How many serious games have you played?

	Never	1 – 5 games	6 – 10 games	Over 10 games
Answers:	22	40	28	7
Total responses:	97			

8. The serious games you played are effective.					
	Strongly disagree	Partly disagree	Neutral	Partly agree	Strongly agree
Answers:	5	13	25	28	10
Total responses:	81				

9. How often do you use speech recognition?					
	Never	Rarely	Sometimes	Often	Very often
Answers:	15	26	33	10	1
Total responses:	85				

10. Why do you use speech recognition? (Multiple choice)				
	Productivity	Curiosity	Entertainment	Other
Answers:	51	32	10	0
Total responses:	93			

11. The speech recognition feature is helpful.					
	Strongly disagree	Partly disagree	Neutral	Partly agree	Strongly agree
Answers:	4	11	25	34	12
Total responses:	86				

12. How much do you enjoy the speech recognition? (Rate 1 to 5)					
Rate:	1	2	3	4	5
Answers:	5	9	31	33	8
Total responses:	86				

The first four questions aimed to discover the background of participants such as age, gender, occupation, and field of study.

The answers from the first question reveal that about 47% of respondents' age are between 15 and 25 years old. Respondents who are between 25 and 35 are around 40%. The rest are between 35 to 44 years old with 11.11%. There is no one who is over 45 years old in the survey.

The responses of the second question shows that female participants in the survey is 43.44% out of 93 participants. Male participants are 49.5% and the rest did not share their gender information.

The third question of the questionnaires asked participants about their employment status. Overall, around 57% of responses were “I’m a student”, 15.16% were “I’m employed person”, 14.14% of answers were “I’m a freelancer”. There is only 2.2% of respondents chose “I’m an unemployed person”.

The next question’s results show the fields in which participants currently studying or studied. It is easy to notice that technology and business are the two fields that participants studied or are studying the most with 35 and 34 people respectively following by natural sciences fields with 13 people. Participants who have knowledge in arts, history, and medicine are the least groups with 9 people in total.

The question 5 to 8 purposed to explore the habit of playing game of participants and how they think about serious games.

When the researcher asked people in the survey how much time they play games, about 40% of respondents answered they spend about 1 to 5 hours a week playing games, 32.33% of people play games for 5 to 15 hours per week, 22.22% of participants spend 15 to 25 hours for games. Only 4.4% of participants play games for over 25 hours per week.

Question 6 purposed to discover the popularity of game types that participants often play. According to the results, multiplayer online games, role-playing games, and action-adventure games are the 3 types of games participants playing the most with 21, 20, and 19 participants playing respectively. This figure is 15 with fighting games, 9 with escape games. Shooting games, sports games, and casual are the least popular types of game with 13 participants in total.

Question 7 was aimed to find how many serious games participants have played. The results show that among 98 people who joined in the survey 40 participants have played 1 to 5 serious games, 28 participants have played 6 to 10 serious games. The number of participants who have played over 10 serious games was only 7. However, there are 22 respondents who have not played any serious game.

Question 8 which was to examine the effectiveness of serious games participants have played. The question asked whether the serious games participants have played were effective, 25 out of 98 responses partially agreed with the question, 10 responses completely agreed, together there were 35. 25 participants did not have a clear opinion. The number of people who strongly disagreed and partly agreed with the question is 18 in total.

The last four questions aimed to investigate participants' opinion about speech recognition.

Question 9 was to examine how often participants use speech recognition technology. The results display that about 33% of 85 participants sometimes use speech recognition, 25.3% of participants rarely use it and 15.18% have not used it before.

The research continues with question 10 that purposed to explore the main reason they use speech recognition. The results illustrates the reasons behind the participants' motivation of using speech recognition. It is the fact that productivity is the main reason with 50 answers followed by curiosity with about 30 responses. There are only 10 responses with entertainment reason.

Question 11 was to examine participants' opinions on how helpful speech recognition is. in general, there 38 respondents thought the speech recognition system is helpful. 25 out of 86 participants did not give their opinions and there are 15 respondents had bad opinions on the helpfulness of the speech recognition.

Question 12 was to examine how much people in the survey enjoy the speech recognition system. According to the results, most participants rated the speech recognition grade 4 and grade 3 with 33 and 31 participants respectively. The number of participants who gave the highest grade is 5. The speech recognition feature received grade 2 and 1 from 14 participants in total.

Questionnaire 2 results:

	Strongly disagree	Partly disagree	Neutral	Partly agree	Strongly agree
1. The speech recognition feature is easy to use.	0	1	4	10	6
2. You feel natural, human-like conversation through speech recognition feature.	0	1	8	11	1
3. The speech recognition helps you concentrate.	0	0	4	12	5
4. The speech recognition responds quickly.	2	13	5	1	0
5. The speech recognition understands your words.	1	6	9	5	0
6. How much do you enjoy hand-free interaction when playing? (Rate 1 to 5)	1	1	5	10	4
7. The speech recognition is boring	1	10	8	4	0
8. Speech recognition should be implemented more on serious games.	1	2	6	10	3

The first question was created to determine the usability of the speech recognition feature in the tested game. Question 1 asked whether testers agree or not with the statement: the speech recognition feature is easy to use. The results reveals the answers of testers. It is a fact that the answer of “partly agree” has the highest number of answers with 10 responses following by “strongly agree” answers with 6 responses. There are 4 answers with the neutral opinion, 1 answer with a partially disagreeing opinion.

Question 2 of the second survey aimed to discover the participants of the survey about the conversation between the player and the example game whether the conversation is human-like or not. 12 participants thought that the conversation is natural the same as

human talking with a human. 8 people did not agree nor disagree. There is only one participant who did not agree with other participants.

The target of question 3 was to explore how speech recognition technology helps players to concentrate. 18 out of 22 participants had more concentration by using speech recognition. No participant had any concentration while using the speech recognition.

Question 4 was designed to examine how quickly speech recognition reacts. The research found that 15 participants did not think the feature work fast as they expected, 5 participants had no opinion and only one participant thought the speech recognition reacted quickly as shown in figure 10.

Question 5 asked the participants whether the speech recognition feature from the example game understood their spoken words. It is clear that the number of participants who did not give their points of view is the largest group with 9 out of 22 participants. The number of people who did not agree with the statement from the question is slightly higher than the ones who agreed with 7 and 5 participants respectively.

The aim of question 6 was to evaluate the voice-user-interface of the speech recognition system in a serious game from players' views. Players were asked to grade hand-free interaction of the game from 1 to 5 where 1 is bad and 5 is very good. According to the results, the number of people gave grade 4 and 5 are 14 participants in total, 5 participants gave grade 3. Two participants considered the hand-free interaction grade 1 and 2.

Question 7 asked participants' opinion on the statement: "the speech recognition is boring". The goal of the question was to evaluate the users' satisfaction. The results show that 10 participants partly disagreed with the statement, 1 participant strongly disagreed. There are 8 neutral answers, 4 answers are "partly agree".

Question 8 was to determine how participants thought of installing more speech recognition on serious games. The results demonstrate that 3 out of 22 players strongly agreed that speech recognition should be implemented more on serious games, 10 players partly agreed while 6 respondents had a neutral opinion and 3 people disagreed.

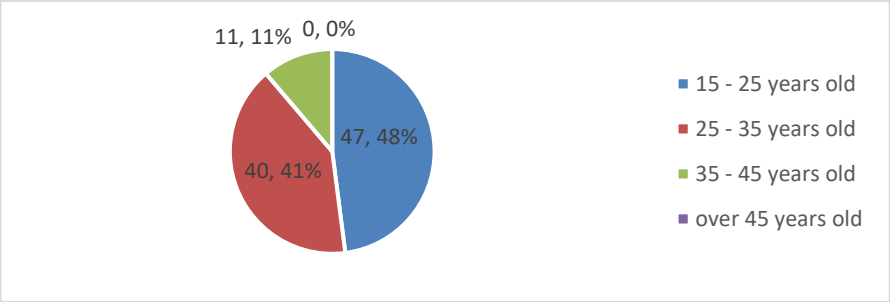

Figure 6. Results of question 1 – Questionnaire 1

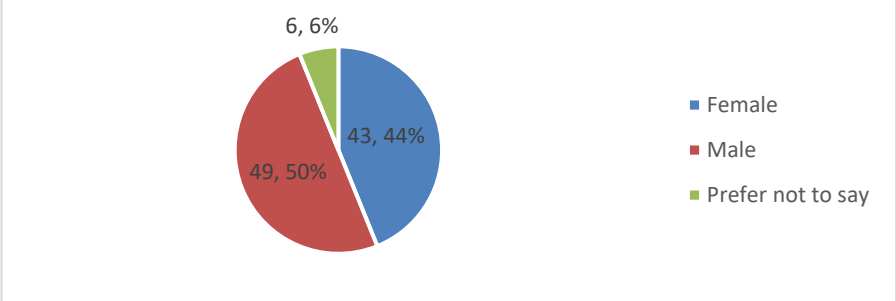

Figure 7. Results of question 2 – Questionnaire 1

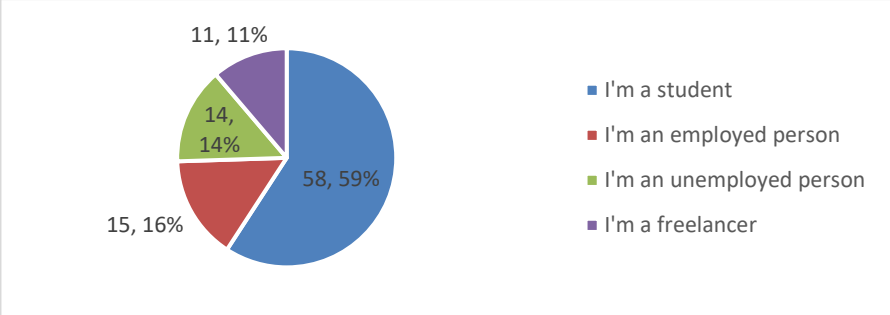

Figure 8. Results of question 3 – Questionnaire 1

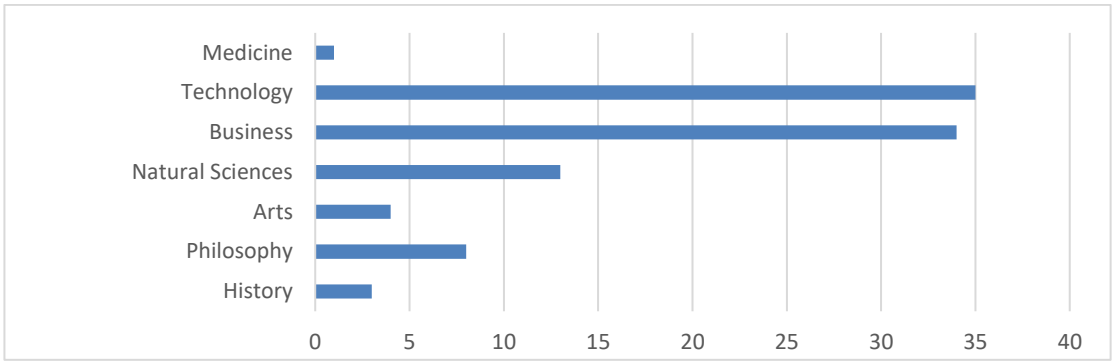

Figure 9. Results of question 4 – Questionnaire 1

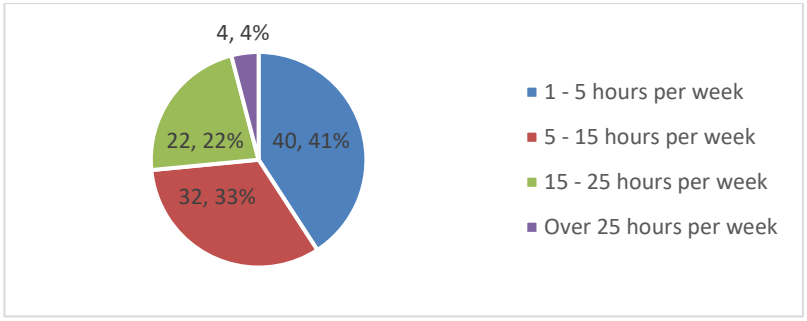

Figure 10. Results of question 5 – Questionnaire 1

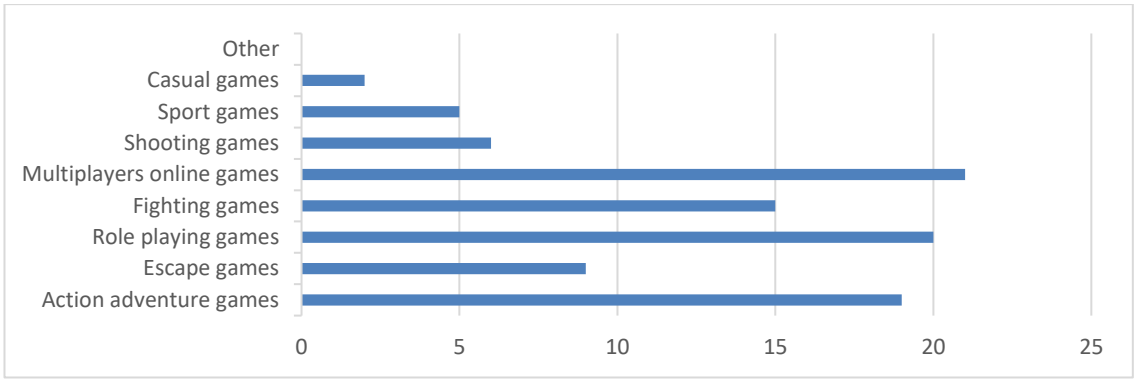

Figure 11. Results of question 6 – Questionnaire 1

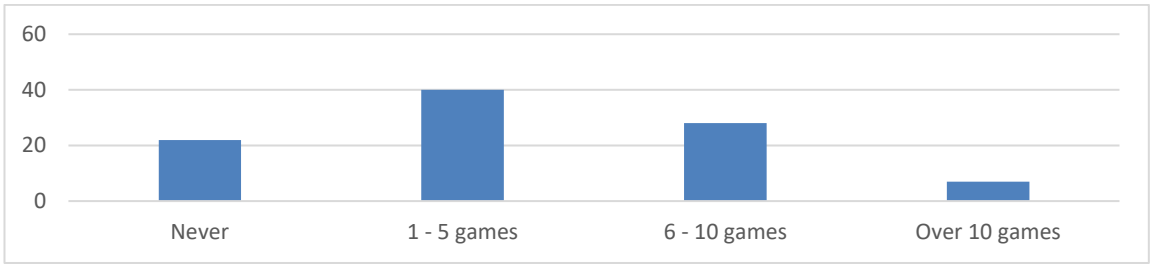

Figure 12. Results of question 7 – Questionnaire 1

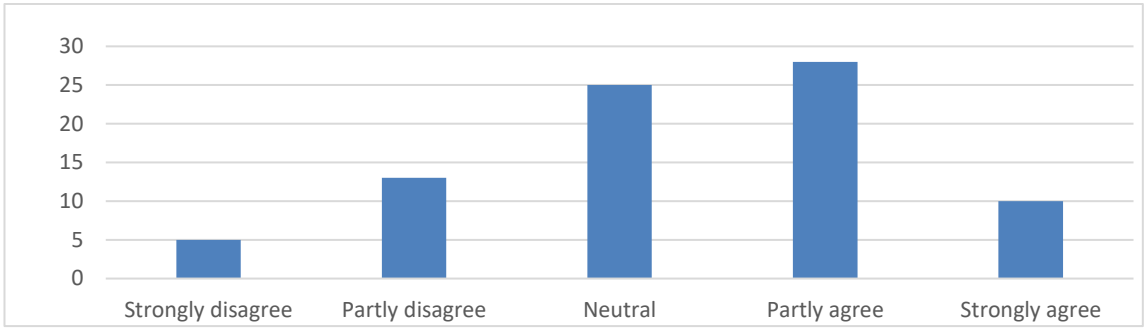

Figure 13. Results of question 8 – Questionnaire 1

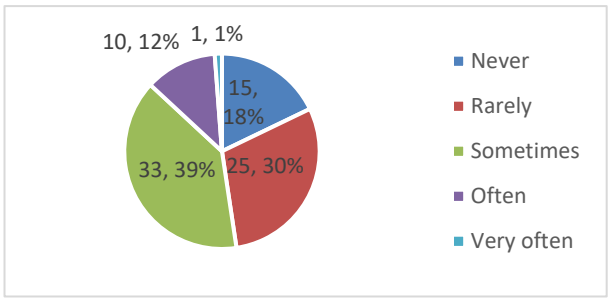

Figure 14. Results of question 9 – Questionnaire 1

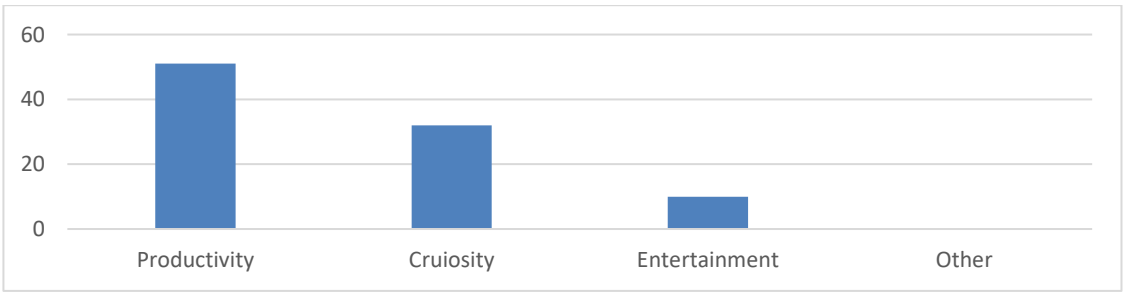

Figure 15. Results of question 10 – Questionnaire 1

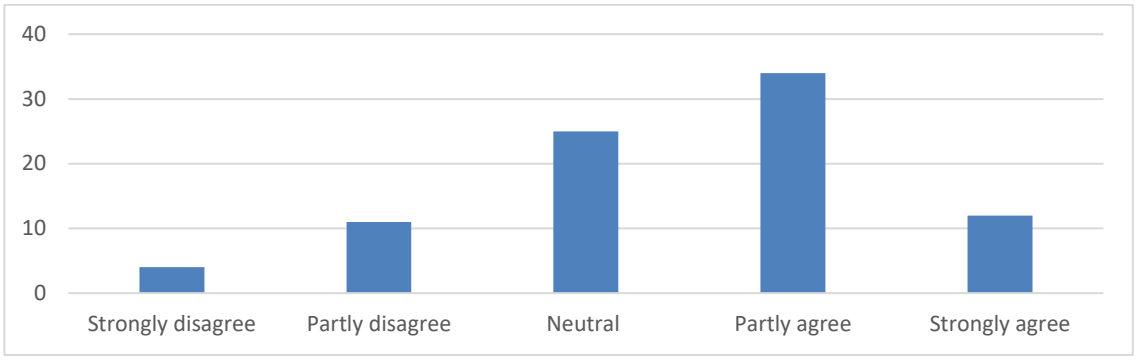

Figure 16. Results of question 11 – Questionnaire 1

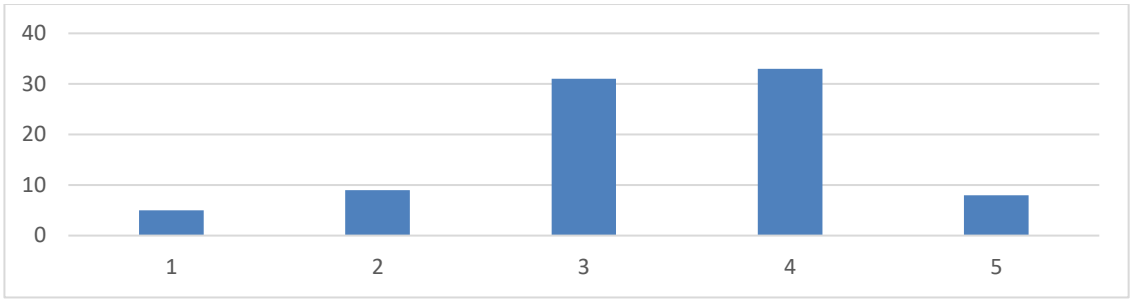

Figure 17. Results of question 12 – Questionnaire 1

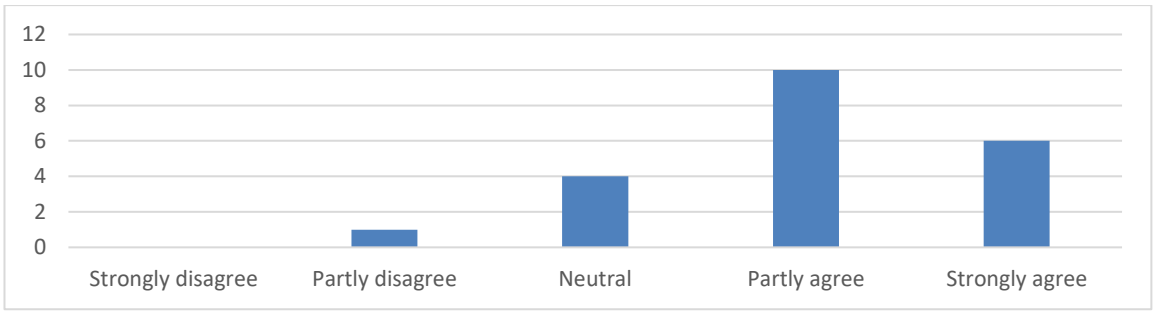

Figure 18. Results of question 1 – Questionnaire 2

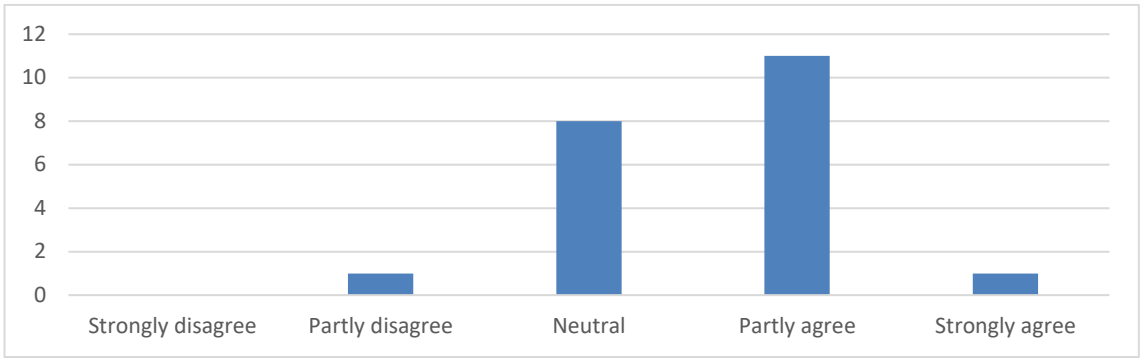

Figure 19. Results of question 2 – Questionnaire 2

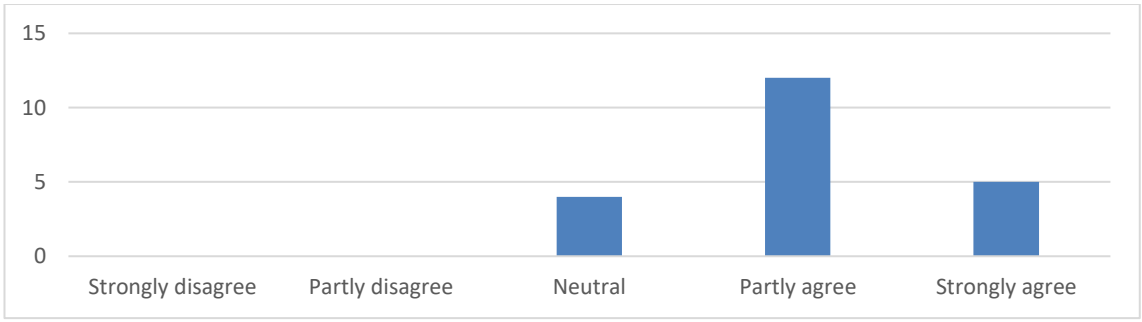

Figure 20. Results of question 3 – Questionnaire 2

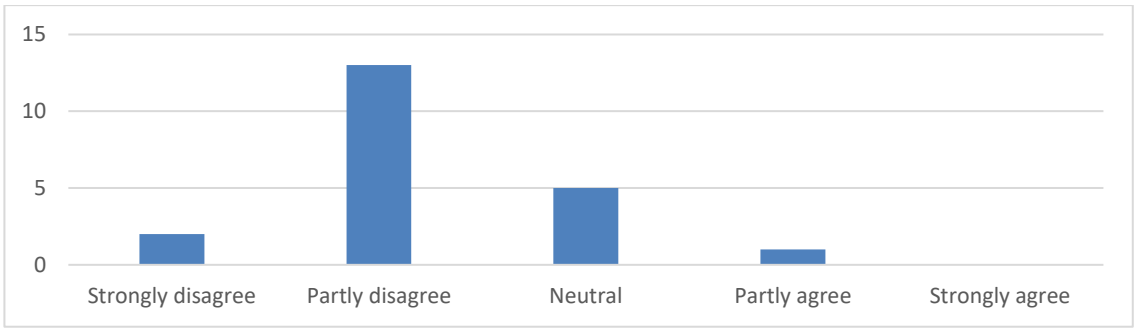

Figure 21. Results of question 4 – Questionnaire 2

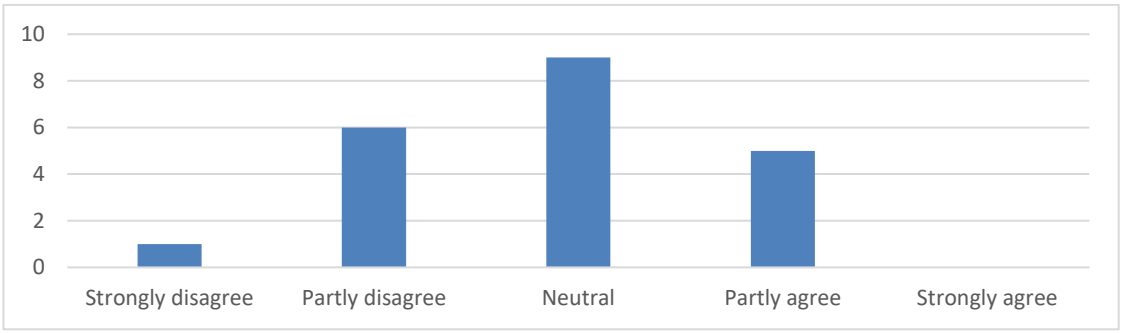

Figure 22. Results of question 5 – Questionnaire 2

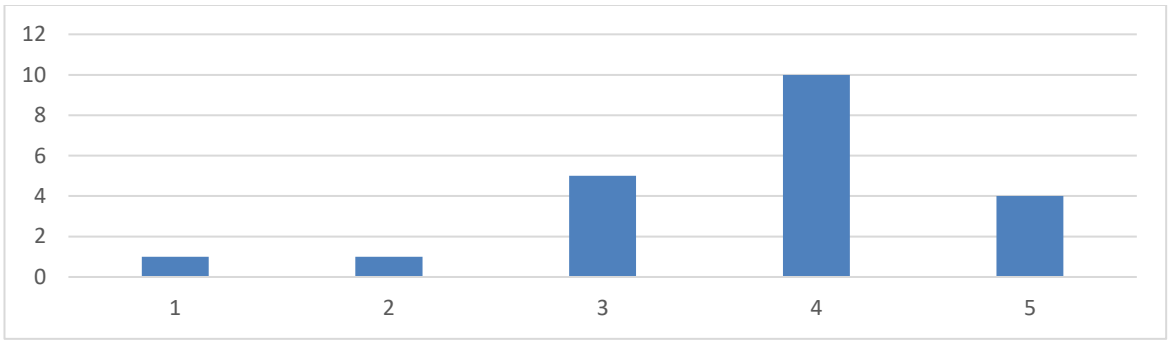

Figure 23. Results of question 6 – Questionnaire 2

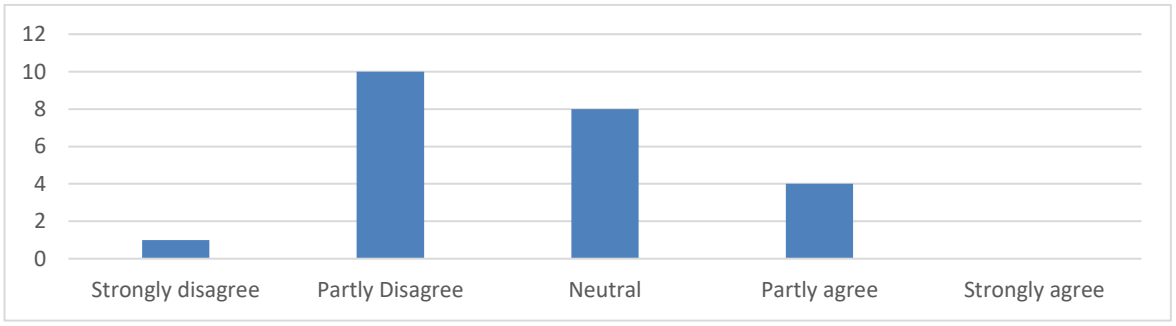

Figure 24. Results of question 7 – Questionnaire 2

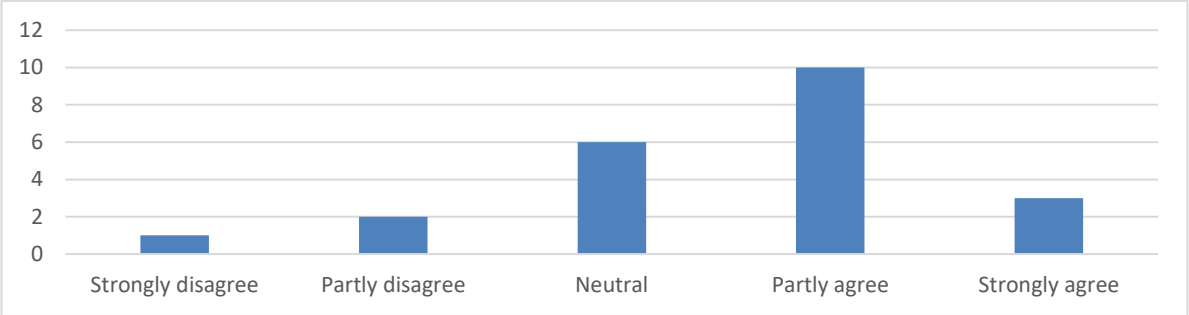

Figure 25. Results of question 8 – Questionnaire 2