Users' Modality Selection in Multimodal System

Min Chul Cha

The Graduate School
Yonsei University
Department of Industrial Engineering

Users' Modality Selection in Multimodal System

A Dissertation
Submitted to the Department of Industrial Engineering
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Min Chul Cha

December 2022

This certifies that the dissertation
of Min Chul Cha is approved.

Thesis Supervisor: Yong Gu Ji

Thesis Committee Member: Wooju Kim

Thesis Committee Member: Kwangwon Ahn

Thesis Committee Member: Seul Chan Lee

Thesis Committee Member: Sol Hee Yoon

The Graduate School
Yonsei University
December 2022

LIST OF CONTENTS

List of Contents ..... **i**

List of Tables ..... **v**

List of Figures ..... **vii**

Abstract ..... **ix**

CHAPTER I: INTRODUCTION 1

1. INTRODUCTION 1

1.1. Background and Problem Statement 1

1.2. Organization of this Dissertation 5

CHAPTER II: BACKGROUND LITERATURE 7

2. LITERATURE REVIEW 7

2.1. Voice User Interface 8

2.2. Modality Selection 9

2.3. Interaction Efforts of Modality 11

2.4. Satisfaction 14

2.5. Interaction Unit of Voice and Touch 14

2.6. Menu structure 16

2.7. User Context 17

CHAPTER III: STUDY 1 19

3. RESEARCH MODEL 19

4. OBJECTIVE AND HYPOTHESES 21

4.1. Objective 21

4.2. Hypotheses 22

5. METHODOLOGY 25

5.1. Experimental Program 25

5.2. Modality Selection Task 26

5.2.1. Voice modality 27

5.2.2. Nonhierarchical modality selection task 28

5.2.3. Hierarchical modality selection task 29

5.2.4. Material for Voice Modality 30

5.3. Pilot Study 31

5.3.1. Task Design 32

5.3.2. Participants 33

5.3.3. Result 34

5.3.4. Pilot Summary 37

5.4. Variables 38

5.4.1. Independent Variables 38

5.4.2. Dependent Variables 39

5.5. Participants 42

5.6. Apparatus and Settings 42

5.7. Procedure 44

5.8. Data Collection and Analysis 45

6. RESULTS 47

6.1. Modality Usage 47
6.2. Interaction Efforts and Satisfaction 53
6.2.1. Interaction efforts and satisfaction of touch 53
6.2.2. Interaction efforts and satisfaction of voice 58
6.2.3. Interaction efforts and satisfaction between menu structure and modality 65
6.2.4. Touch modality efforts between menu structures 71
6.3. Modality Switching Points 72
7. DISCUSSION 80

CHAPTER IV: STUDY 2 84
8. RESEARCH MODEL 84
9. OBJECTIVE AND HYPOTHESES 86
9.1. Objective 86
9.2. Hypotheses 87
10. METHODOLOGY 89
10.1. Context 89
10.2. Variables 91
10.2.1. Independent Variables 91
10.2.2. Dependent Variables 92
10.3. Participants 92
10.4. Apparatus and Settings 93
10.5. Procedure 93
10.6. Reproducibility of Study 95

11. RESULTS.....98
11.1. Modality Selection..... 98
11.1.1. Non-hierarchy..... 98
11.1.2. Hierarchy..... 102
11.2. Physical Effort..... 106
11.3. Mental Effort..... 108
11.4. Satisfaction..... 111
12. DISCUSSION 113

CHAPTER V: GENERAL CONCLUSION 117
13. OVERALL SUMMARY 117
13.1. Summary of Study 1..... 117
13.2. Summary of Study 2..... 118
14. CONCLUSION 119
14.1. Conclusion and recommendations..... 119
14.2. Contribution..... 121
14.3. Limitation and Future studies..... 122

REFERENCES 124

APPENDIX 140

ABSTRACT (IN KOREAN).....150

List of Tables

Table 2.1. User contexts classified by physical and mental resources 18
Table 5.1. Descriptive statistical analysis of pilot study 34
Table 5.2. Result of t-test of one character text entry..... 35
Table 5.3. Independent Variables of Study 1 38
Table 5.4. Dependent variables 40
Table 6.1. Descriptive statistics of voice and touch usage 48
Table 6.2. Summary of logistic regression model of the modality usage..... 50
Table 6.3. Result of ANOVA of physical effort of non-hierarchy touch 54
Table 6.4. Result of ANOVA of mental effort of non-hierarchy touch 54
Table 6.5. Result of ANOVA of satisfaction of non-hierarchy touch..... 54
Table 6.6. Result of ANOVA of physical effort of hierarchy touch 56
Table 6.7. Result of ANOVA of mental effort of hierarchy touch..... 56
Table 6.8. Result of ANOVA of satisfaction of hierarchy touch 57
Table 6.9. Result of three-way ANOVA of Physical Effort of voice modality..... 59
Table 6.10. Result of three-way ANOVA of Mental Effort of voice modality..... 61
Table 6.11. Result of three-way ANOVA of Satisfaction of voice modality..... 63
Table 6.12. Result of two-way ANOVA for physical effort by modality and menu structure..... 65
Table 6.13. Simple main effect of menu structure on physical effort in both modalities.. 66
Table 6.14. Result of two-way ANOVA for mental effort by modality and menu structure 67
Table 6.15. Simple main effect of menu structure on mental effort in both modalities 67
Table 6.16. Result of two-way ANOVA for satisfaction by modality and menu structure 69
Table 6.17. Simple main effect of menu structure on satisfaction in both modalities..... 69
Table 6.18 T-test results of physical effort and mental effort of touch modality between menu structures. 71

Table 6.19. results of ANOVA for modality on interaction efforts and satisfaction in both menu structures 73
Table 6.20. Result of Games-Howell post-hoc analysis on physical effort..... 75
Table 6.21. Result of Games-Howell post-hoc analysis on mental effort 77
Table 6.22. Result of Games-Howell post-hoc analysis on satisfaction 79
Table 10.1. Independent Variables of study 2 92
Table 10.2. Result of logistic regression analysis with study 1 and 2 as IV 96
Table 11.1. Descriptive statistics of voice and touch usage in non-hierarchy..... 99
Table 11.2. Summary of logistic regression model of the modality usage in non-hierarchy. 100
Table 11.3. Descriptive statistics of voice and touch usage in hierarchy 103
Table 11.4. Summary of logistic regression model of the modality usage in hierarchy.. 104
Table 11.5. Result of three-way ANOVA for physical effort by menu structure, context, and modality..... 106
Table 11.6. Result of three-way ANOVA for mental effort by menu structure, context, and modality 109
Table 11.7. Result of three-way ANOVA for satisfaction by menu structure, context, and modality 112

List of Figures

Figure 1.1. U.S. Smart Speaker Frequency of Use 2
Figure 1.2. Average command counts over 12 weeks (Cho et al., 2019) 3
Figure 1.3. Daily Voice Assistants Use Case Frequency in Vehicle (Voicebot, 2020) 4
Figure 2.1. Sample of SMEQ survey 13
Figure 2.2. An example of the composition of words and syllables of the Korean..... 16
Figure 3.1. The research model of the study 1 ..... - 20 -
Figure 5.1. Screenshot of the first page of program 26
Figure 5.2. Stimuli of non-hierarchy (a) and hierarchy (b) conditions 27
Figure 5.3. Speech recognition system for voice modality of this study 28
Figure 5.4. Examples of non-hierarchy modality selection task 29
Figure 5.5. Examples of hierarchy modality selection task 30
Figure 5.6. Snapshot of text entry task for pilot study 32
Figure 5.7. Snapshot of simple keyboard touch task for pilot study 33
Figure 5.8. The voice usage by experimental conditions in simple keyboard touch..... 36
Figure 5.9. Snapshot of survey page 41
Figure 5.10. (a) Experimental setting of study 1 and (b) example scene of experiment... 43
Figure 5.11. Overall procedure of study 1 45
Figure 6.1. Probability of voice usage predicted by the logistic regression model..... 51
Figure 6.2. ROC curve of the logistic regression model of study 1 52
Figure 6.3. Interaction efforts and satisfaction of touch modality depending on number of touches in non-hierarchy condition 55
Figure 6.4. Interaction efforts and satisfaction of touch modality depending on the number of touches in hierarchy condition 57
Figure 6.5. Physical effort of voice modality according to number of touches and syllables per touch 60

Figure 6.6. Mental effort of voice modality according to number of touches and syllables per touch..... 62
Figure 6.7. Satisfaction of voice modality according to number of touches and syllables per touch..... 64
Figure 6.8. Difference in physical effort by modality according to menu structure 66
Figure 6.9. Difference in mental effort by modality according to menu structure..... 68
Figure 6.10. Difference in satisfaction by modality according to menu structure 70
Figure 6.11. Physical effort by modality in each menu structure..... 75
Figure 6.12. Mental effort by modality in each menu structure..... 77
Figure 6.13. Satisfaction by modality in each menu structure 79
Figure 8.1. The research model of the study 2 85
Figure 10.1. Experimental setting for four contexts..... 90
Figure 10.2. Overall procedure of study 2 94
Figure 10.3. Probability of voice usage predicted by logistic regression model with study 2..... 97
Figure 11.1. Predicted probability of voice usage by contexts in non-hierarchy 101
Figure 11.2. Predicted probability of voice usage by contexts in hierarchy 105
Figure 11.3. Differences in physical effort by modality according to context in each menu structure..... 107
Figure 11.4. Differences in mental effort by modality according to context in each menu structure..... 109
Figure 11.5. Differences in satisfaction by modality according to context in each menu structure..... 112

Abstract

Users' Modality Selection in Multimodal System

Min Chul Cha

Department of Industrial Engineering

The Graduate School

Yonsei University

Many smart devices being developed are equipped with voice personal assistants or voice-based interactive agents that can command by voice. Users can perform multiple functions to achieve their goals by choosing the traditional interfaces or Voice User Interface (VUI). Although voice interfaces have various advantages such as eyes- and hands-free, intuitiveness, and high input speed, VUI usage is gradually decreasing, and users only use voice interaction in a few domains. This is because voice interaction was not selected in competition through repetitive use. Therefore, it is necessary to have a high understanding of the modalities suitable for each function and context and to study the

features of modalities selected by users through the comparison between modalities in a multimodal system.

This dissertation aims to compare the user's modality selection in multimodal systems using modality interaction units and to analyze and evaluate the effects of modality features, menu structure, and context of use. To present a design guide for modalities in multimodal systems, this dissertation would achieve the following research goals.

- To analyze and evaluate the effect of modality features and menu structure on modality selection, interaction efforts, and satisfaction (Study 1).
- To identify differences in modality selection, interaction efforts, and satisfaction according to the context of using the multimodal system (Study 2).

To achieve the goal of study 1, a multimodal system that presents voice and touch modalities under various conditions was developed, and an experiment was conducted to repeat the modality selection. The interaction units of touch and voice modalities were defined as one touch and one syllable, respectively, and the number of touches (1~5) and the syllable per touch (1~ 8) were used as design variables. In addition, the type of function was defined as a design variable called menu structure (non-hierarchical and hierarchical). In the experiment of study 1, the modalities selected by users, physical and mental efforts required for interaction of each modality, and satisfaction with each modality were evaluated depending on the differences in design variables.

In study 1, the results showed that the user's modality selection was significantly influenced by syllable per touch and menu structure. As a reference point for the

interpretation of modality selection, the point at which voice modality usage exceeds 50% was defined as the modality switching point. The modality switching points occurred at 2~3 syllable per touch in non-hierarchy, and at 4~ 5 syllable per touch in hierarchy. However, there was no significant difference according to number of touches. In addition, it was confirmed that the design variables affected the physical & mental efforts, and satisfaction of modalities, and based on this, the user's modality selection was made.

In study 2, the context of using multimodal systems was classified by physical and mental resources, and a total of four contexts (baseline, watching, reading, and driving) were defined. The modality selection task was conducted under the contexts, and it was tested whether the contexts, along with previous design variables, influenced the user's modality selection and subjective factors (physical effort, mental effort, and satisfaction).

As a result of Study 2, the user's modality selection moved toward using voice more according to the contexts. The modality switching points moved to a higher syllable per touch in the order of baseline < watching< reading < driving. That is, the context of using physical resources increased voice modality usage more than the context of using mental resources, and an interaction effect of the two resources was also found. The subjective factors of modalities were also influenced by contexts.

In this study, it was confirmed that users' modality selection in a multimodal system is determined by the ratio of interaction units between modalities, which varies depending on the type of function or context of use.

In this dissertation, multimodal systems were evaluated by an interaction unit-based approach in the fields of ergonomics and HCI. Based on the results of this dissertation, the design factors and guidelines of voice modality were provided to improve users' voice modality use in a multimodal system. These findings have significance as basic research that can be applied to design that enhances the usability of the entire system through the advantages of voice modality characteristics such as high performance and low workload.

Keywords: Multimodal System, VUI, Modality Selection, Interaction Effort

CHAPTER I: INTRODUCTION

1. INTRODUCTION

1.1. Background and Problem Statement

Recently, people have encountered smart devices supporting personal assistant services everywhere. Besides smartphones and smart speakers, in-vehicle infotainment and the smart home services that manage home devices are rapidly permeating our daily lives. These devices provide not only the traditional graphical user interface but also provide a modality option which is a voice- or speech-based interface. The market for voice-based systems is expected to grow quickly, with the number of voice assistants will reach 8.4 billion by 2024 (Vailshery, 2021).

Due to the multimodal system provided by voice-based devices, people became to be able to select one of several options to achieve their goals. They can perform tasks using desired or useful modalities depending on the type of task or context. In particular, Voice User Interface (VUI) has some advantages such as eyes- and hands-free, intuitiveness, and fast input speed (Pearl, 2016). Because of these advantages, VUI shows higher performant

and gives lower cognitive and physical load than the traditional modalities, and it allows users to operate devices without any problems in not only the single task situation but also multitasking situations (Cherubini et al., 2009; A. L. Cox et al., 2008; J. Kim et al., 2019; Tsimhoni et al., 2004)

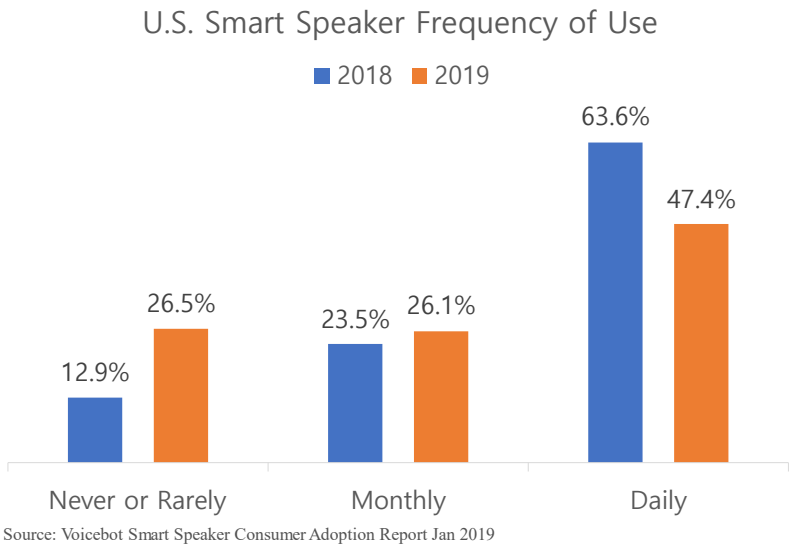


Figure 1.1. U.S. Smart Speaker Frequency of Use

However, although functions that can be performed in voice assistants are increasing year by year, the rate of using VUI is not increasing anymore. The number of skills registered in the U.S. Amazon Alexa store exceeded 80,000 in 2021 from 56,000 in 2019 (Voicebot, 2021). Nevertheless, the percentage of daily VUI use is declining (Voicebot, 2019) (see Figure 1.1), and even in 2022, the total number of VUI users in the U.S.

decreased from 132 million to 123.5 million. In several studies, it was also observed that users gradually stopped using VUI despite its advantages. Cho et al. (2019) conducted an observational study on the behavior of users using conversational agents at home. In that study, users explored their agents with interest for the first 1-2 weeks, but after that, they used only minimal functions or no longer used agents (see Figure 1.2). It seems that users no longer use VUI due to repetitive tasks and experiences. Therefore, it was necessary to find out what factors caused users to stop using voice modality and returned to the conventional method.

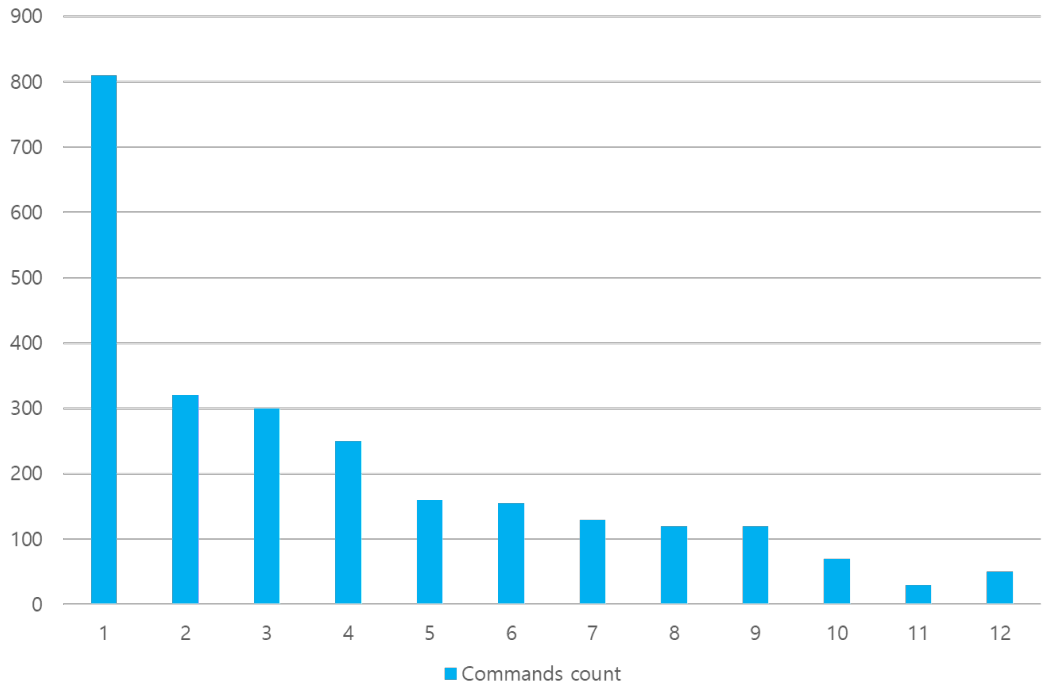


Figure 1.2. Average command counts over 12 weeks (Cho et al., 2019)

The frequency of voice modality uses was also very diverse depending on the type of functions or the contexts of use. Bentley and colleagues (2018) observed the use of smart speaker assistants in a home over a long period using Google Home devices. They found that users most often use smart speakers for listening to music, with usage exceeding 40%. Next, they used smart speakers for searching for information, setting an alarm, and checking the weather. There was a difference in the usage rate of voice modality by functions, and the usage in the vehicle showed another aspect. As shown in Figure 1.3, voice assistants in vehicles were often used to make a phone call, send a text, or ask for directions. However, it did not mean that users stopped using the functions they used a lot at home. This indicates that there are additional functions performed by voice as the context of use changes. Therefore, it was necessary to study the usage of voice modality by users according to the features of the function to be developed with voice or the context of use.

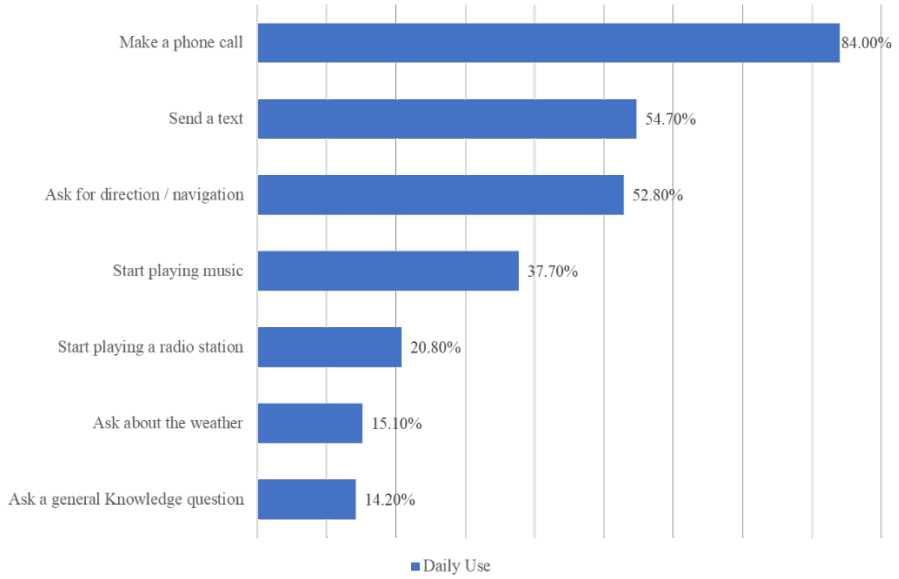


Figure 1.3. Daily Voice Assistants Use Case Frequency in Vehicle (Voicebot, 2020)

This dissertation aimed to figure out in which modality conditions users use or do not use voice modality under various contexts of use and features of functions to be performed. In order to explore this objective, a comparison was conducted in a multimodal system that provided the touch modality and voice modality together. The research consisted of two studies, and each study dealt with the effects of modality features and menu structure, and the effect of usage context, respectively. Although previous studies were limited to specific functions, this study has a unique significance in that it suggested the universal design factors of voice interface to enhance users' usage of voice modality through modality features and menu structure in a multimodal system. Thus, it is expected that the result of this study can be a study applied to various menu structures and modality features in real systems.

1.2. Organization of this Dissertation

This dissertation is composed of five chapters and fourteen sections.

CHAPTER I: Introduction

CHAPTER I addresses the limitation of existing studies as the background and the motivation of study. Based on this, the problem statement of this study is presented. Also, it describes the explanations on the organization of the dissertation.

CHAPTER II: Background Literature

CHAPTER II reviews the literature related to voice user interface, modality selection,

interaction efforts, satisfaction, menu structure, and user context. Based on this literature review, each section provides several hints on the independent variables and the dependent variables and task design that will compose study 1 and study 2.

CHAPTER III: Study 1

In CHAPTER III, the experiment was conducted to examine the effects of modality features and menu structure on modality selection in multimodal systems. Section 3 & 4 present the research model, objective, research question, and hypotheses of study 1. Section 5 explains the methodology of study 1 including the experimental program, tasks, variables, participants, apparatus/settings, procedure, and data collection/analysis. The results and discussion of the experiment are presented in sections 6 and 7.

CHAPTER IV: Study 2

In CHAPTER IV, the experiment was conducted to investigate the effects of contexts on the user's modality selection in multimodal systems. Likewise, the research model, research purpose, and hypothesis are described from section 8 to section 9. Section 10 explains the overall methodology including context. The results and discussion of the study 2 are presented in section 11 and 12.

CHAPTER V: General conclusion

CHAPTER V provides the overall summary of both studies and conclusions of this dissertation. Also, it presents several recommendations for designing voice modality based on the finds of this dissertation.

CHAPTER II: BACKGROUND LITERATURE

2. LITERATURE REVIEW

This chapter will go over the review of the Voice User Interface, and the latest technologies and advantages studied to enhance the usability of multimodal systems (section 2.1). Next, the topic is related to the process of user modality selection in a multimodal system that will be examined to identify the causes that are not used even in the advantages of VUI and the benefits of performance (section 2.2). The third part describes physical effort, mental effort, and satisfaction perceived by the operating modalities, and relates to how to measure each variable (section 2.3 & section 2.4). The fifth part reviews the interaction units of each modality for designing variables for comparisons between modalities and explains how they will be used as modality features in the study (section 2.5). The sixth part is about the menu structure, another independent variable used for task design in studies 1 & 2, that will be examined the characteristics of the multimodal system according to the hierarchy of system (section 2.6). The seventh part includes the context and classification criteria in which multimodal systems are used in various multi-tasking situations and will be used in study 2 (section 2.7).

2.1. Voice User Interface

Voice user interface is a conversational interface in which a user interacts with a system via speech (Perlman et al., 2019; Strayer et al., 2019). These conversational interfaces include prompts, grammar, and dialog logic (Cohen et al., 2004). In this way, VUI allows users to verbally access information and services through interaction in everyday language. In addition, compared to the existing interfaces (e.g., touch, physical buttons, etc.), it has various advantages, such as freedom of eyes and hands, intuitiveness, and fast input speed mentioned before (Pearl, 2016). Although voice recognition technology is not perfect, it is evolving day by day, and it is expected that it will eventually develop to the level of understanding users' natural speech (Saon et al., 2021).

Due to the development of voice recognition technology and the advantages of VUI itself, VUI is installed in various systems and competes with existing modalities. Many studies showed that VUI is superior to other interfaces in terms of performance or workload (J. Hong & Findlater, 2018; Mitchard & Winkles, 2002; Perlman et al., 2019; Ruan et al., 2016). In addition, voice has a feature that allows input by skipping several steps of the interface, such as a shortcut key, so that multiple touches can be replaced with one voice, reducing physical load (Perugini et al., 2007).

Nevertheless, users' voice usage is still relatively low, and many researchers are struggling to address it (Beirl et al., 2019; Porcheron et al., 2018; Sciuto et al., 2018; van Pinxteren et al., 2020). Researchers had developed several techniques for VUI (‘open-mike’

or ‘barge-in’), error recovery methods, etc., but they have not become fundamental solutions (J. Kim et al., 2019; Mane et al., 1996; Perakakis & Potamianos, 2008a) Schaffer et al. (2011) approached this problem using modality selection in mobile systems. Therefore, focusing on this approach, this study intended to investigate the usage of voice and touch modalities considering more in-depth modality features.

2.2. Modality Selection

Most smart devices are multimodal systems that provide more than one modality option. Users repeatedly perform the same task using various modalities and find the optimal modality for the task n those multimodal systems (Cherubini et al., 2009; Suhm et al., 1999). The process of finding such an optimal modality is called modality selection, and it was revealed that the efficiency and effectiveness of interfaces affects the user's modality selection (Bilici et al., 2000; Schaffer et al., 2011).

The effectiveness is related to the accuracy of interfaces (Card et al., 1990; Chen & Tremaine, 2006). For the interface to be highly effective, the system must understand the user's commands accurately and return accurate results. This is an issue that can be solved by the development of interface implementation technology, and the existing touch or physical input devices is almost equivalent to error-free (Chen & Tremaine, 2006). As voice recognition technology advances, voice interfaces are also moving toward error-free

systems (J. Li, 2022). Therefore, it is believed that effectiveness will be improved in the future and is no longer within the scope of this study.

On the other hand, the efficiency refers to how quickly or easily users can achieve their goal using an interface (Perakakis & Potamianos, 2008b). In terms of input speed, voice is known to have a faster input speed than other modalities. When a typing expert uses a keyboard, one of the traditional interfaces, the typing rate is 80 wpm (Mitchard & Winkles, 2002). However, it has been found that the speech rate of normal people is about 200 wpm and is much faster than keyboards. Furthermore, Ruan et al. (2016) found that the VUI is about 3 times faster than the touch keyboard. VUI also enables users to easily achieve their objectives. Unlike existing GUI-based interfaces, voice does not require eyes or hands, so it is relatively physically free (Du et al., 2018; J. Hong & Findlater, 2018). It also enables users to jump to the final step without performing tasks sequentially (Jung et al., 2020; Perugini et al., 2007; Resnick & Virzi, 1992).

In the following studies, a multimodal system that could operate with touch or voice was presented to users. The researchers expected that the efficiency of each modality would affect the users' modality selection, that is which modality they use to perform a task. By controlling the number of touches and syllables per touch (S/T), the efficiency level of each modality was controlled and the changes in users' modality selection were observed. Through this, this study aimed to find the modality switching point, the point at which voice modality starts to be used more than touch.

2.3. Interaction Efforts of Modality

The effort required for interaction with a certain modality can affect the expected utility, cost, or benefit of interfaces, which can even influence the users' modality selection (Gray et al., 2006; Parasuraman et al., 2000). Budiu (2013) defined 'interaction cost' as the sum of physical and mental effort that users allocate efficiently to achieve their goals. She said that users try to maximize the utility of their behaviors by reducing the interaction cost and tend to find a modality to match it. Therefore, evaluating the effort required to interact with modalities can be a hint for which modalities users will use to perform tasks.

Measuring the effort required for users to interact with interfaces is one of the most important research topics in the HCI field. The effort required for users to interact with interfaces has been evaluated and managed in terms of physical or mental aspects (Cabral et al., 2005; Liu & Thomas, 2017; Schwaller & Lalanne, 2013). And each effort is evaluated in either objective or subjective ways on various devices.

Cha et al. (2017) used surface electromyography (EMG) to measure thumb muscle activity while using smartphones. It was found that the activity of muscles is very different depending on the form factor of the smartphone. Although this measurement method is a very objective indicator, it has some difficulties due to complicated measurement methods and equipment attached to users. Therefore, in this study, we decided to use the subjective

method of directly asking users about their interaction efforts.

The subjective measure involves the self-reporting measure in which participants report their perceived physical and mental workload to interact with a system. NASA-TLX is one of the most popular multidimensional scales in this perspective (Arrabito et al., 2015; Hwangbo et al., 2013; Laureiti et al., 2017; Turner et al., 2020). It consisted of a survey about the mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). However, since NASA-TLX is an indicator that covers many dimensions, measurements tend to be complex and limit the number of observations and samples. Therefore, it is not suitable for this study, which needs to evaluate the effort that varies by experimental conditions in both modalities.

For simple evaluation, the literature review was conducted to find a one-question questionnaire. The Subjective Mental Effort Questionnaire (SMEQ) is a unidimensional survey used to measure subjective mental effort (Zijlstra, 1993). It is based on a 16-point scale from 0 to 150 with nine labels from “Absolutely no effort” to “Extreme effort” (See Figure 2.1) (Park et al., 2021; Widyanti et al., 2013). Sauro & Dumas (2009) revealed that SMEQ performed best among the single questionnaires. It was easy to learn to use, had high correlations with other measures, and was highly sensitive. Therefore, in this study, we decided to evaluate the interaction efforts of each modality using SMEQ and modified the questions to physical and mental efforts (Figure 5.9).

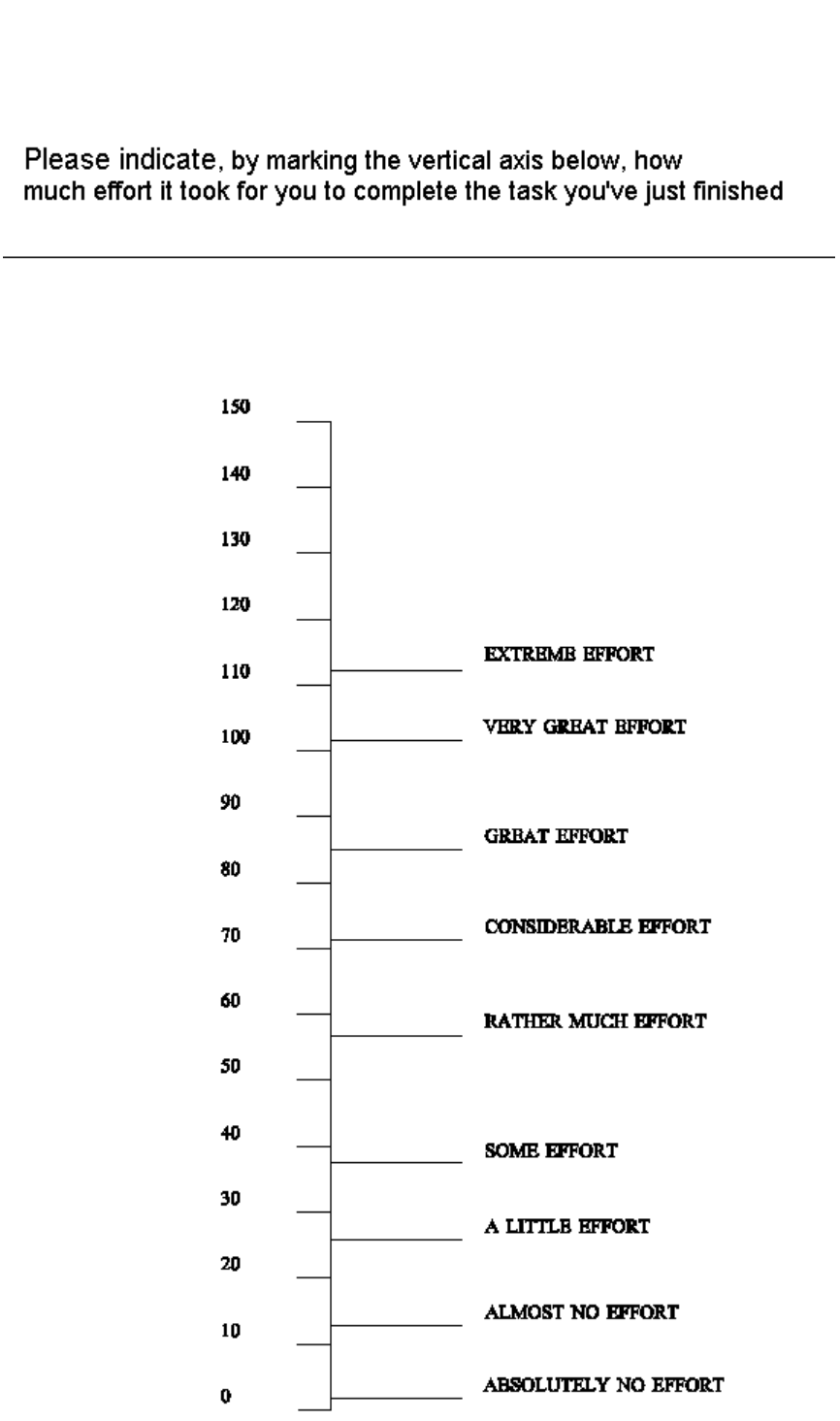


Figure 2.1. Sample of SMEQ survey

2.4. Satisfaction

Satisfaction as well as effectiveness and efficiency are considered as one axis of usability. ISO 9241-11 defined usability as: “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (Standardization, 1998). It was assumed that changes in modality features would also affect satisfaction, and an additional questionnaire was used to evaluate satisfaction.

As with other indicators, satisfaction was also assessed through a single questionnaire. Based on the evaluation method used by Westbrook (1980), this study also evaluated the satisfaction of voice and touch modality on a scale of 0 to 100 points.

2.5. Interaction Unit of Voice and Touch

In order to evaluate interaction effort and satisfaction and compare users' modality choices, it was necessary to adjust the difficulty of each modality. In this study, the interaction units of touch and voice modalities were changed to change the interaction effort of each modality. A literature review was conducted to define the interaction unit of touch and voice modality.

In the case of touch, analysis of operation units has already been conducted in many studies. Based on the keystroke level modeling proposed by Card et al. (1980), many studies have taken one touch as the unit of touch interaction (Schaffer et al., 2011). Lee et al. (2019) conducted a study to investigate task completion time and task performance in a vehicle using this touch unit.

While the interaction unit of physical buttons or touch is obvious, the unit of voice is complex and ambiguous. In particular, there are few studies defining this in terms of human-computer interaction. Schaffer and his colleagues (2011) used one-touch as an interaction step of touch modality in a mobile environment. However, they did not set up an interaction unit for voice modality and defined all voice interactions as having one interaction step. This might be a very different approach from how people perceive speech. People perceive the units of speech in three ways: phonemes, syllables, and words (Wickens et al., 2021). However, from the viewpoint of speech production or speech recognition research, it has been studied that the unit of speech production is a syllable (Fujimura, 1975; Wijnen, 1988). Therefore, in this study, the interaction unit of voice for input was defined as a syllable, which is a unit of speech production.

The form of syllables is defined by the different rules according to various languages. Especially in ‘Korean’, characters consist of a combination block of up to four graphemic letters, and such a character served as a syllable (Simpson & Kang, 2004) (see Figure 2.2). Therefore, for developing a Korean multimodal system, the length of characters was

regarded as the length of syllables, and it served as the interaction unit of voice in this study.

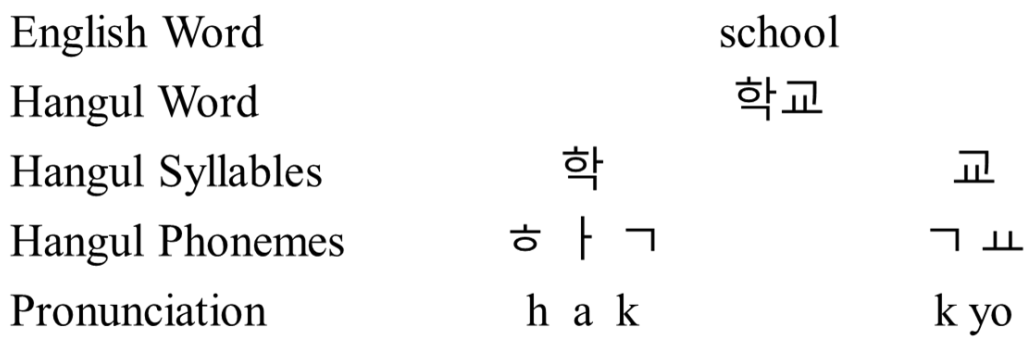


Figure 2.2. An example of the composition of words and syllables of the Korean

2.6. Menu structure

The structure of the menu affects the behavior and mind of users executing the function. When users use the system, they decide what action to take next, which is affected the menu structure. This is because what information needs to be remembered and what information needs to be newly explored will change. There are two menu structures, one is a non-hierarchical structure (Christie et al., 2004; Han & Kwahk, 1994; Paap & Cooke, 1997) and the other is a hierarchical structure (Jiang & Chen, 2022; R. Li et al., 2017).

In the actual system, two types of menu structures (non-hierarchy & hierarchy) consist of complex tasks, and, when users performed each task, there are various patterns of spending physical and mental efforts. The non-hierarchical tasks require repetitive physical

efforts without switching displayed items, such as volume control and radio channel change (Han & Kwahk, 1994). On the other hand, the 'hierarchical' task is a task in which physical action and item searching are repeatedly performed, and it requires repetitive mental effort from users (Medhi et al., 2013). When using the GUI, the menu structure of the system can affect the physical and mental efforts for operating modality, but when using the VUI, commands can be input regardless of the menu structure. These differences can cause changes in the relative efforts of modalities by task, which consequently affect the modality selection. Therefore, it is expected that there would be a difference in the usage of voice and touch interface by these two menu structures, and this study aims to verify it.

2.7. User Context

In interacting with the interface, the user's situation is very important in selecting a specific modality to operate the interface. Depending on the type or usage of the interactive channel being used in various usage situations of the user, the working memory required for the user to use the new modality is limited (Oviatt et al., 2004; Wickens, 2002). In addition, multiple tasks accessing the same cognitive resources cause a high workload and decrease work performance. In order to avoid an increase in workload, it is common for users to select a relatively free interaction channel in a specific situation (Lemmelä, 2008) In previous studies, contexts were analyzed using human capabilities (hand, eyes, ears, voice, and working memory) or some types of workloads (aural, visual, physical, and

cognitive) to analyze voice- and eye-based tasks in different situations. were classified (Jameson & Kl¨ockner, 2005; Lemmelä et al., 2008).

In this study, based on the existing classification method, the contexts were classified according to the availability of human physical and mental resources. Accordingly, in this study, four contexts were created: baseline, reading, watching, and driving, and each resource used is shown in Table 2.1 below.

Table 2.1. User contexts classified by physical and mental resources

User Context		Physical Resource	
		Low	High
Mental Resource	Low		
		Baseline	Reading
	High		
		Watching	Driving

CHAPTER III: STUDY 1

3. RESEARCH MODEL

Study 1 aimed at investigating the effects of modality features and menu structure on the user's modality selection in multimodal systems. And this also aimed to figure out how those features affect the physical effort, mental effort, and satisfaction of voice and touch modalities. Finally, by analyzing how interaction efforts and satisfaction and modality choices were related, this study tried to reveal their mediating effects between modality features, menu structure and modality selection. The research model of the study 1 is presented in Figure 3.1.

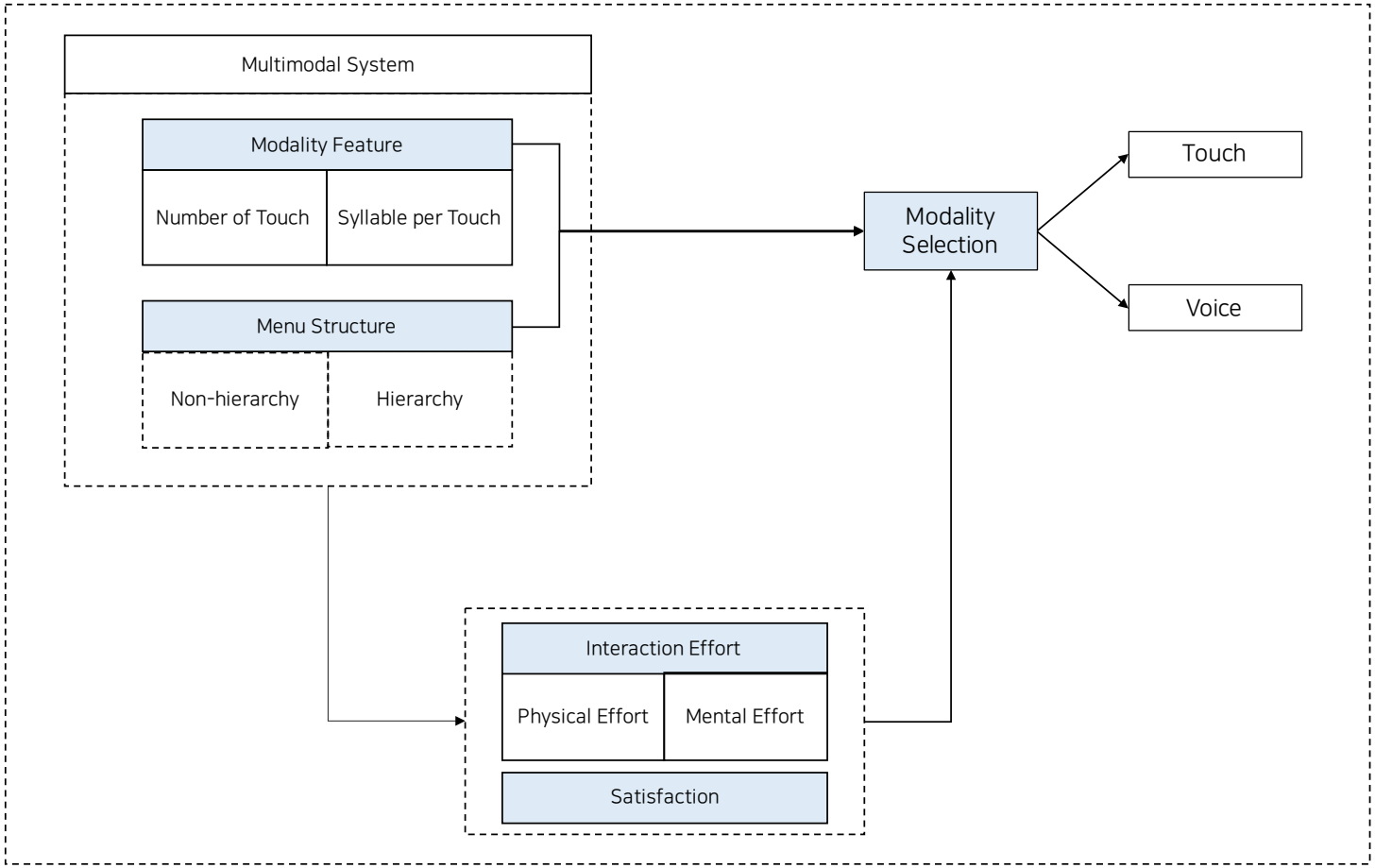


Figure 3.1. The research model of the study 1

4. OBJECTIVE AND HYPOTHESES

4.1. Objective

As mentioned in Section 3, the object of this study was to examine the effects of modality features and menu structure on modality selection in multimodal systems, and the following research questions were presented below:

- RQ1: How will modality features and menu structure affect the users' modality selection?
- RQ2: How will modality features and menu structure affect interaction efforts and satisfaction of each modality?
- RQ3: When will modality switching occur?
- RQ4: How will the usual VUI usage affect modality selection?

To answer the given research questions and achieve the research objective of this study, this modality selection study developed and tested the hypotheses. Each hypothesis is for determining if the independent variables of the multimodal system change the user's modality selection, interaction efforts, and satisfaction of voice and touch modality.

4.2. Hypotheses

Hypothesis 1: In a multimodal system, the changes in the independent variables would affect the user's modality selection.

H1a: The usage of voice modality would be increase as the number of touches increases.

H1b: The usage of voice modality would be increase as the syllable per touch decreases.

H1c: The usage of voice modality in hierarchical task would be higher than thar of non-hierarchical task.

The number or length of items to be entered affect the difficulty of manipulation. This is because the number of manipulations of the modality changes, which can influence the time or effort required for the task. However, in both touch and voice modalities, which are to be compared in this study, the difficulty of tasks would increase as the number of manipulation units increases. Lamel et al. (2002) and Naumann et al. (2008) argued that as the length of input attribute increased, voice modality was preferred. On the other hand, Perakakis & Potamianos (2008a) reported that users did not use voice modality for "long" attributes. However, since the voice rate is basically faster than the touch rate (Cherubini et al., 2009), as the length of the task increases, the difference in the amount of input per unit time (e.g., words per minute (WPM)) would accumulate and the difference in time or effort required to complete the task would also increase. Therefore, it was expected that voice modality usage would increase as the number of touches increased. On the other hand,

syllable per touch (S/T) is a factor that affects only the voice modality without changing the touch modality. Therefore, as S/T decreases, voice usage was expected to increase.

In a hierarchical menu structure, users usually need to perform additional item searching. For this reason, many studies have reported that the hierarchical menu structure is more difficult than the non-hierarchical menu structure (Christie et al., 2004; Han & Kwahk, 1994; Wallace et al., 1987; Zaphiris et al., 2002). However, voice modality is not dissipative from the information structure and could skip several levels of the menu structure (Du et al., 2018). Accordingly, in the voice modality, it was expected that there is no difference in difficulty according to the menu structure, but in the touch modality, an increase in difficulty is expected.

Hypothesis 2: In a multimodal system, the changes in the independent variables would affect physical effort, mental effort, and satisfaction of each modality.

H2a: As the number of touches increases, physical and mental efforts of each modality would increase, and satisfaction of each modality would decrease.

H2b: As the syllable per touch increases, physical and mental efforts of voice modality would increase, and satisfaction of voice modality would decrease.

H2c: Depending on modalities, the effects of menu structure on physical effort, mental effort, and satisfaction would be different (there is an interaction effect of selected modality and menu structure).

Hypothesis 2 was developed based on the explanations described in establishing Hypothesis 1 above. It was expected that increasing the difficulty of each modality would increase the modality's physical and mental effort and decrease satisfaction. Therefore, detailed hypotheses were established to test the modalities of difficulty that were expected to change in Hypothesis 1.

Hypothesis 3: There would be Modality Switching Points at the intersection of the modality's physical effort, mental effort, and satisfaction by modality features and menu structure.

Many studies have been conducted based on the fact that people can perceive the required physical and mental effort and satisfaction according to changes in task difficulty (Peissner & Doebler, 2011; Reimer et al., 2013). Thus, people might be able to decide which modality to use based on each modality's interaction efforts or satisfaction.

Hypothesis 4: The more often people use VUI in their everyday lives, the more they use the voice modality.

Based on the Technology Acceptance Model (TAM), a traditional model in the HCI field, VUI would have formed users' existing usage patterns by various factors (Nguyen et al., 2022; Rupp et al., 2018). It could be possible that people who have high VUI usage are already familiar with VUI and aware of its high usability (Nguyen et al., 2019). Therefore, high VUI usage behavior would increase the voice modality usage rate.

5. METHODOLOGY

5.1. Experimental Program

The experimental program was developed using JavaScript for presenting the modality selection tasks to participants. All progresses of this experiment were performed through the program. The first page of program was a demographic survey page (Figure 5.1). There were some questions about participants' name, age, gender, and experience of VUI. The next page was for setting the test conditions (Appendix 3.2). After selecting the test conditions, the experiment started automatically, and the current level page was displayed. Then participants pressed the next button, the trial ready page was displayed for 2 seconds, and moved to the modality selection task page. In the task page, participants performed tasks by voice or touch modalities. If participants completed six trials in a certain level, the survey page was present for evaluating interaction efforts and satisfaction of each modality. The data were also collected automatically in the database.

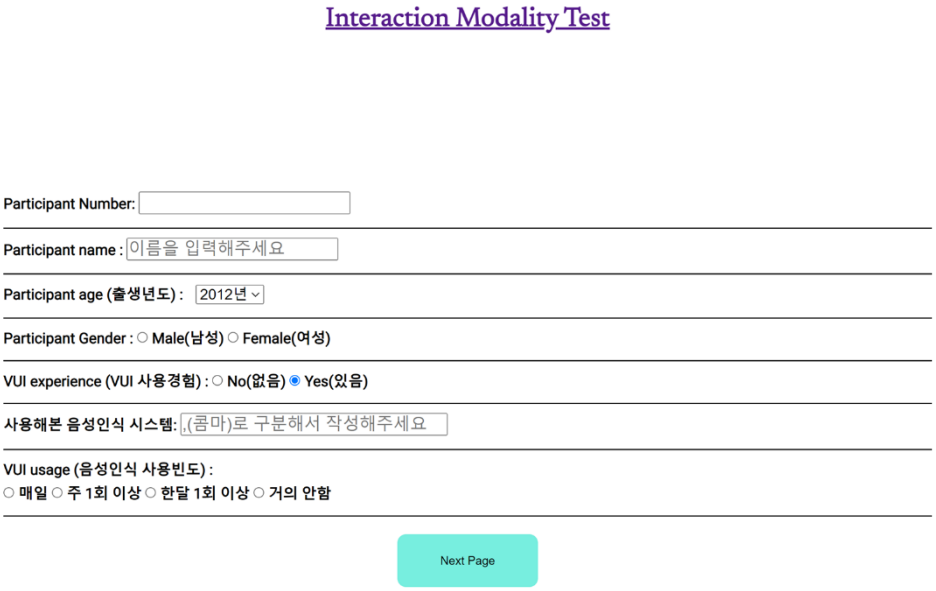


Figure 5.1. Screenshot of the first page of program

5.2. Modality Selection Task

Participants were asked to perform the task of selecting which modality they thought was more efficient to move on to the next step, either touch or voice, on the screen displayed on the touch monitor. The number of touches was controlled from1 to 5, and the length of syllables to speak was varied to 1, 2, 3, 4, 5, 6, 7, and 8 syllables per touch. The buttons for the task were a square size of 30mm x 30mm, and the gap between the buttons is 3mm, which was a level that does not affect touch performance (Hwangbo et al., 2013; H. Kim & Song, 2014) (Figure 5.2).

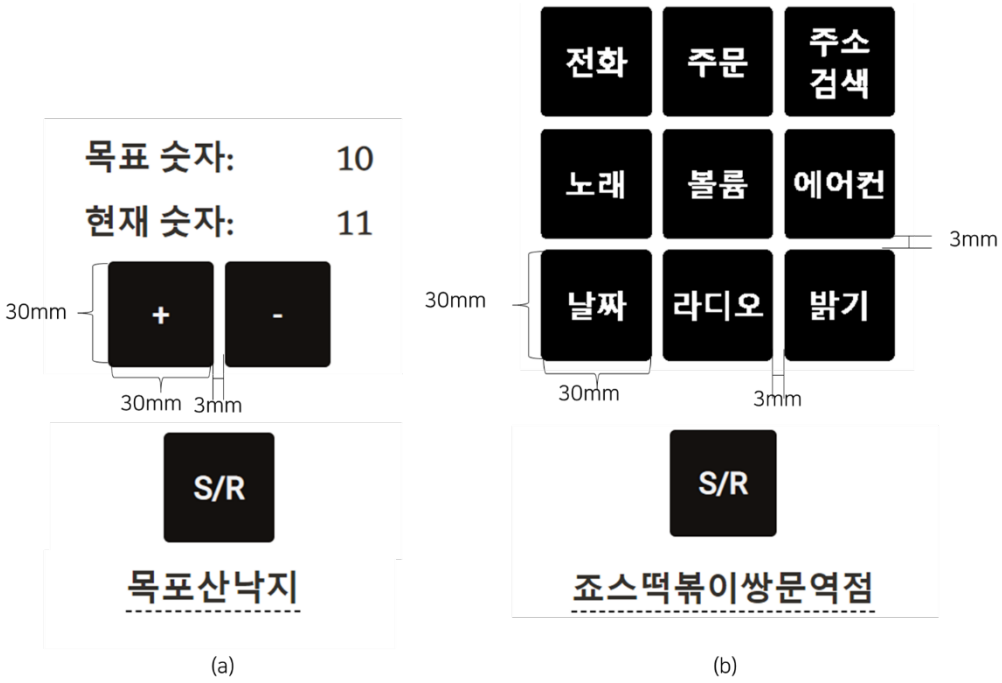


Figure 5.2. Stimuli of non-hierarchy (a) and hierarchy (b) conditions

5.2.1. Voice modality

For voice modality, the task of speaking the name of restaurants via voice recognition was given. One of the restaurant addresses (target length = (number of touches) * (S/T)) was presented as the target of the voice task. If the target length exceeded 20 syllables, two random restaurants were combined and presented. Participants were asked to touch a button written ‘S/R’ to activate the speech recognition engine (Push-to-Talk). When the ‘S/R’ button was touched, the ding-dong sound was used to notify the voice recognition activated,

and the participant spoke the name of the restaurant. Although it was implemented to enable actual voice recognition, this experiment was conducted with fake voice recognition that randomly occurs errors with a 2.5% probability after the participant spoke. This was because this study assumed that errors rarely occur due to improved voice recognition technology, and participants did not recognize that it was fake voice recognition or did not try to speak incorrectly.

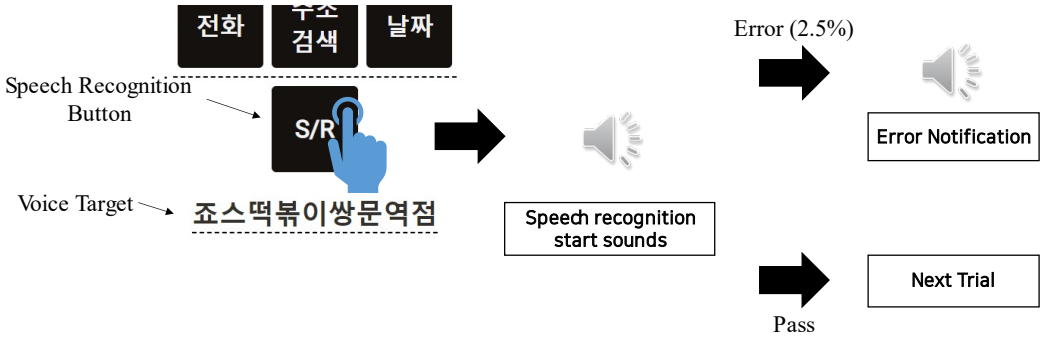


Figure 5.3. Speech recognition system for voice modality of this study

5.2.2. Non-hierarchical modality selection task

In non-hierarchical touch was the task in which a cognitive judgment and repetitive touches are performed, such as tunning volume, temperature, and radio channel. A number between 10 and 90 was randomly presented as a target number. A current number was given as a number randomly added or subtracted by the number of touches from the target number (see Figure 5.4). E.g., if the number of touches is 3 and the target number is 23, the current

number could be 20 or 26. Participants determined whether the current number was bigger or smaller than the target number, and matched the numbers by touching '+' or '-' several times. For the voice task in this condition, syllable per touch was controlled with a total of five levels in 1, 2, 3, 4, and 5 S/T.

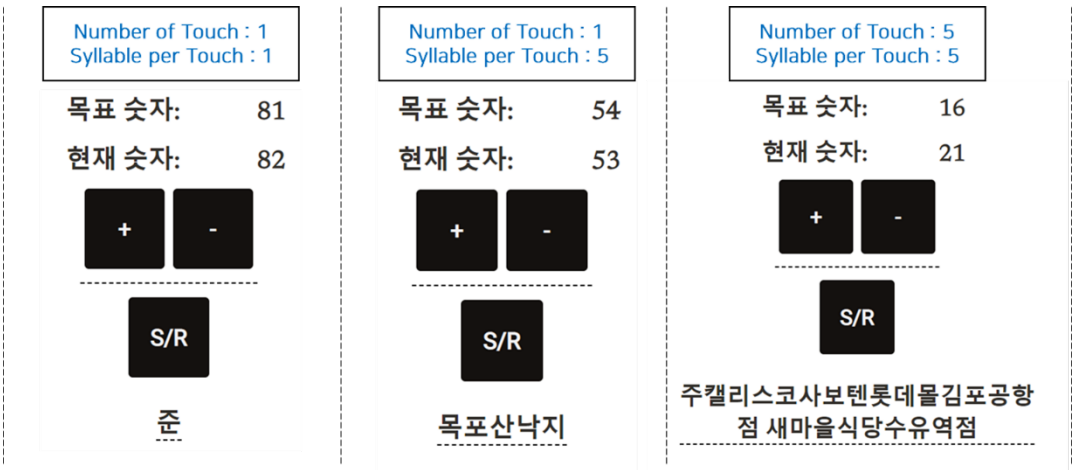


Figure 5.4. Examples of non-hierarchy modality selection task

5.2.3. Hierarchical modality selection task

The hierarchical touch task was designed to test repeatedly exploring the hierarchy of systems, such as categories of web pages. Nine items ('radio', 'music', 'volume', 'date', 'navigation', 'temperature', 'phone', 'order', 'brightness') were randomly placed on nine buttons in the form of 3x3 grid, and one of the nine items was randomly selected as a touch

target (see Figure 5.5). Participants had to touch the target several times depending on the experimental conditions, and the location of the items was randomly rearranged each time they touched it, so they had to search the target for a new arrangement for each touch. For voice tasks corresponding to hierarchical tasks, the syllable per touch was eight levels: 1, 2, 3, 4, 5, 6, 7, and 8 S/T.

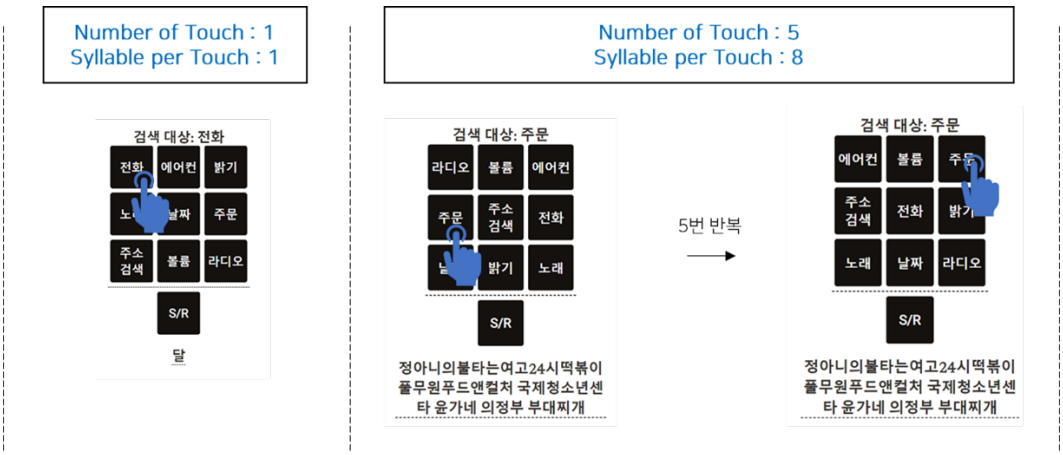


Figure 5.5. Examples of hierarchy modality selection task

5.2.4. Material for Voice Modality

A pilot test was conducted by using 17000 words that were frequently used in Korean, but some words and their random arrangement make participants feel unfamiliar and awkward. To cover this problem, about 480,000 restaurant names registered in local cities

were used as voice materials after removing special characters or numbers. Participants reported that they feel natural to speak a restaurant because it consisted of familiar vocabulary and the name of local regions. From the collected lists, restaurant names with a length matching the number of syllables for each experimental condition were randomly selected and presented to the participants.

5.3. Pilot Study

A pilot test was conducted to design the materials and tasks used in the experiment. In the pilot test, the general vocabulary of the 'Frequent Use in modern Korean Research' announced in 2002 by the National Institute of Korean Language was refined and used as a stimulus (N. H. Cho, 2002). Articles, pronouns, verbs, nouns, adverbs, numerals, and adjectives with more than six frequency counts and less than seven characters were selected from the general vocabulary. A total of 29223 words list were generated. The average syllable per touch of the generated list was 1/2.60, and words of random length were combined to create stimuli with the number of syllables required for the experiment.

The task used was given a text entry task to input a given word and a simple keyboard touch task. A detailed description of this is described in section 5.3.1. The equipment used for the pilot experiment was the same as the equipment used in main experiment.

5.3.1. Task Design

For the pilot test, participants were presented with a modality selection task to choose between touch and voice modalities. First, text entry was a task of inputting words presented randomly through a touch keyboard or voice recognition. Text input was chosen because it is one of the most important and frequent tasks in most devices. The number of characters was increased by one, and words of up to five characters were presented as stimuli. A Qwerty touch keyboard was presented for text entry, and the 'S/R' button was presented along with the keyboard for voice modality (see Figure 5.6). The task was terminated if the subjects performed the tasks using only voices up to three-characters condition.

Figure 5.6. Snapshot of text entry task for pilot study

Next, a simple keyboard touch task was presented. Unlike text entry, this consisted of a blank touch keyboard, and buttons to be touched were randomly colored in green (see

Figure X). Participants were asked to proceed with the tasks by selecting either touching all the green buttons on the blank keyboard or speaking the words presented. The number of touches was set to 2~5 touches, and the syllable per touch was set to 1~4 for the task.

Figure 5.7. Snapshot of simple keyboard touch task for pilot study

In the experiment, text entry and simple keyboard touch tasks were conducted in random order. In simple keyboard touch task, number of touches increased by 1 from 2 touches, and syllable per touch decreased by 1 from 4 syllables. A total of 5 trials were performed for each condition, and 25 modality selection tasks were performed for the text entry and 80 modality selection tasks for the simple keyboard touch.

5.3.2. Participants

Twelve participants (6 males and 6 females) were recruited from the local university. The age of participants ranged from 25 to 37 years (mean age = 30.25 years, SD = 4.24).

The screening questions concerned the participant's age and voice user interface use. The study took approximately an hour to complete. Participants were compensated about $15 for their participation in this study.

5.3.3. Result

5.3.3.1. Text Entry

In the text entry task, the voice modality was overwhelmingly used from one character condition. According to Table 5.1, even when the number of characters is one, the rate of using voice was high at 73.8%. As a result of T-Test to test if it differs from 50%, it was found that the rate of voice modality usage in one character was significantly higher than 50% (p=.000) (see Table 5.2).

Table 5.1. Descriptive statistical analysis of pilot study

Modality	**Number of characters** *				
	1	2	3	4	5
Touch	16 (26.2%)	2 (3.6%)	2 (3.9%)	1 (6.3%)	0 (0%)
Voice	45 (73.8%)	54 (96.4%)	49 (96.1%)	15 (93.8%)	6 (100%)
Total	61	56	51	16	6

*It is as same as syllables in Korean.

Table 5.2. Result of t-test of one character text entry

test value = 0.5	t	*df*	Sig (2-tailed)	Mean Difference	Lower	Upper
Modality	4.186	60	.000	.238	.12	.35

5.3.3.2. Simple keyboard touch

The use of voice modality increased as syllable per touch decreased in simple keyboard touch tasks. As shown in Figure 5.8, it was observed that the voice modality usage exceeded 50% at nearly 2 syllable per touch. However, since the presented stimulus was a random sequence of words, problems such as awkward word order and unfamiliarity of words have been reported by participants.

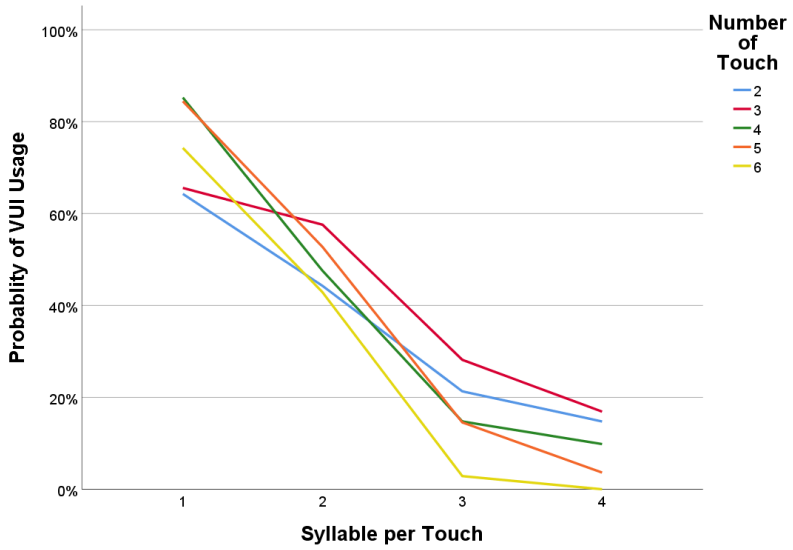


Figure 5.8. The voice usage by experimental conditions in simple keyboard touch

5.3.4. Pilot Summary

This pilot study was conducted for two purposes. First, it was determined that the stimuli were not suitable in terms of the appropriateness of the stimulus. The task was not difficult enough to break the experiment, but as participants reported, the sequence of randomly generated words made it difficult for them to speak naturally. It was because the structure of the sentence was awkward, such as the appearance of words that did not match or words of the same category appearing in a row. Although frequently used words in Korean were extracted, participants often felt that the words were unfamiliar. Therefore, in this study, it was necessary to find other stimuli that are more familiar and natural.

Second, text entry was excluded from the main experiment in terms of the appropriateness of task design. As a result of two types of tasks, it was found that voice was overwhelmingly efficient in the text entry task. In addition, for simple keyboard touch, the point where the voice usage exceeded 50% was located close to two syllables per touch. According to this result, since the Korean words list planned to be used in this study has an average number of 1/2.6 syllables per touch, it was obvious that voice modality already prevailed over touch modality in the text entry condition.

Therefore, in Study 1 and 2, it was decided to design a new type of task, and non-hierarchy and hierarchy tasks which are practical types of touch rather than text entry were designed and tested. This made it possible to compare touch and voice close to the actual functions used in existing systems.

5.4. Variables

5.4.1. Independent Variables

There were three independent variables: (1) menu structure, (2) number of touches, and (3) syllable per touch (S/T) (Table 5.3). Menu structure refers to the shape or the structure of the multimodal systems. It is a characteristic of function that can be performed by voice. Number of touches refers to the number of touches required to be performed in touch modality. Syllable per touch refers to the number of syllables in the voice corresponding to a single touch. The number of syllables to be presented for the voice modality is determined by multiplying the number of touches by syllables per touch, e.g., 3 touches × 2 S/T = 6 syllables.

Table 5.3. Independent Variables of Study 1

Independent Variable	**Level**	
Menu Structure	Non-hierarchy	Hierarchy
Number of Touches	1,2,3,4,5	
Syllable per Touch (S/T)	1, 2, 3, 4, 5	1, 2, 3, 4, 5, 6, 7, 8

5.4.2. Dependent Variables

There were four dependent variables: one objective variable and three subjective variables (see Table 5.4). The modality that participants used to perform the tasks was recorded and used as the objective variable. There were physical effort and mental effort as two subjective variables, and they were items that modified SMEQ into physical and mental questionnaires for this experiment. Participants were asked to evaluate the use of the physical resource (moving hand, making sounds, etc.) and mental resources that are necessary to complete the trial with a specific modality. Both efforts were evaluated separately for each modality, and each questionnaire was a 16-point scale ranging from 0 to 150. The last one was the satisfaction of modality. Participants were asked to score their satisfaction with performing each modality and the questionnaire was an 11-point scale between 0 to 100. Figure 5.9 shows a snapshot of the survey page with questionnaire for each modality.

Table 5.4. Dependent variables

Measures	Description	Scale and Question
Selected Modality	Modality used by participants to perform each task	• Touch(0), Speech(1) • If participants used voice modality to perform all trials in a certain S/T, it would be considered the modality switching
Mental Effort	Mental effort required to complete the task using each modality (*SMEQ)	Q: How much mental effort does it take to perform task using each modality? A: 150 points scale (10 point intervals)
Physical Effort	Physical effort required to complete the task using each modality (*SMEQ)	Q: How much physical effort does it take to perform task using each modality? A: 150 points scale (10 point intervals)
Satisfaction	Satisfaction of each modality to perform the task	Q: How satisfied was each modality to complete the task? A: 100 points scale (5 point intervals)

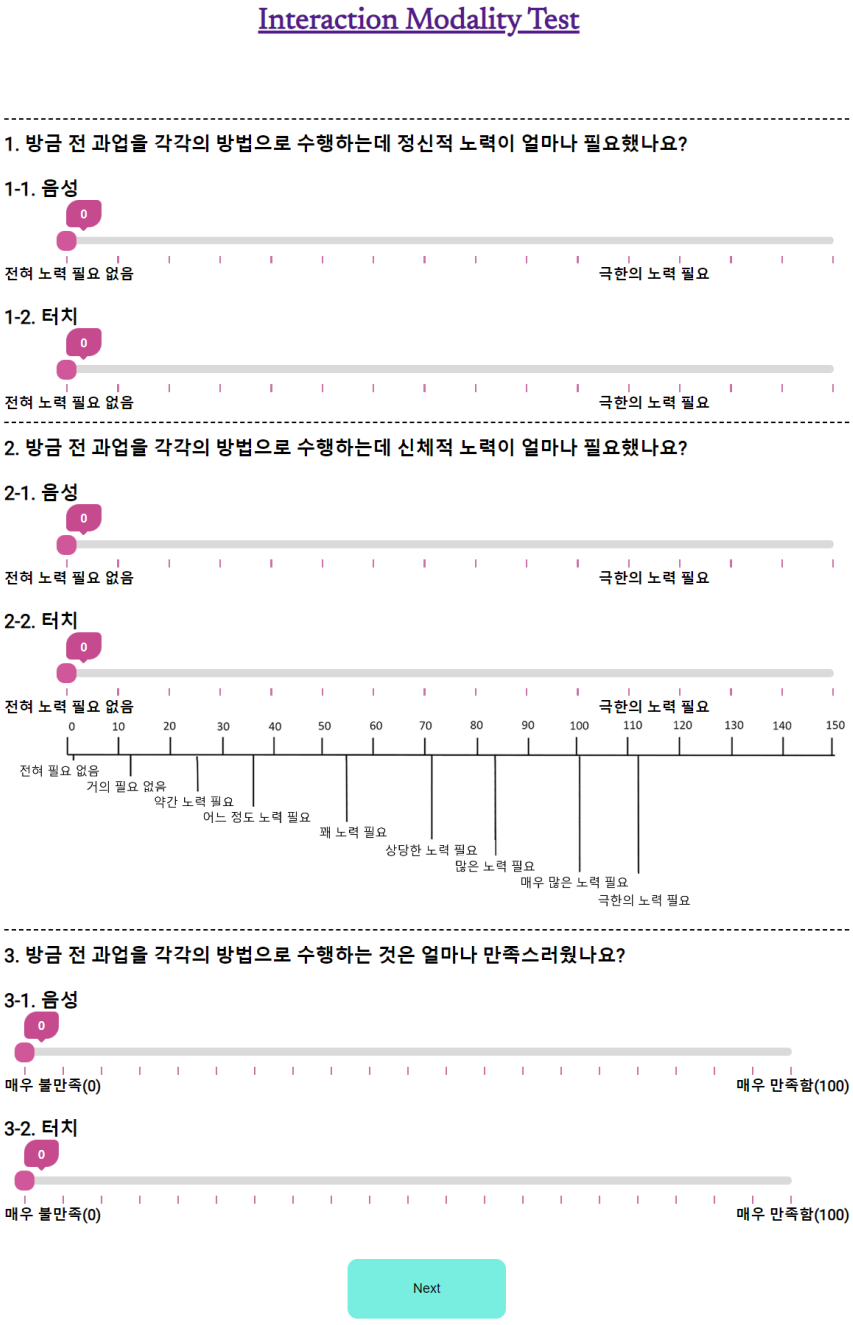


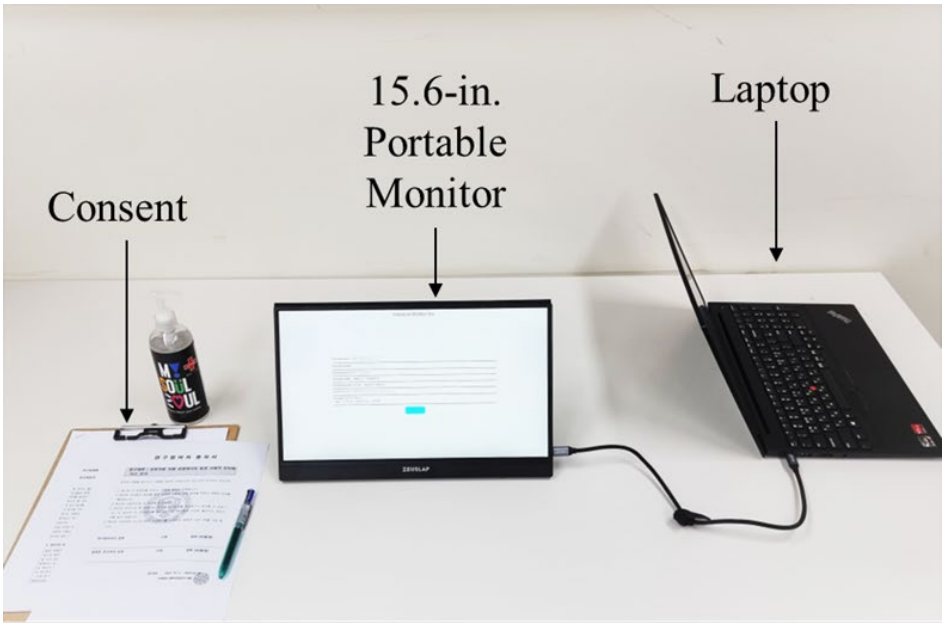
Figure 5.9. Snapshot of survey page

5.5. Participants

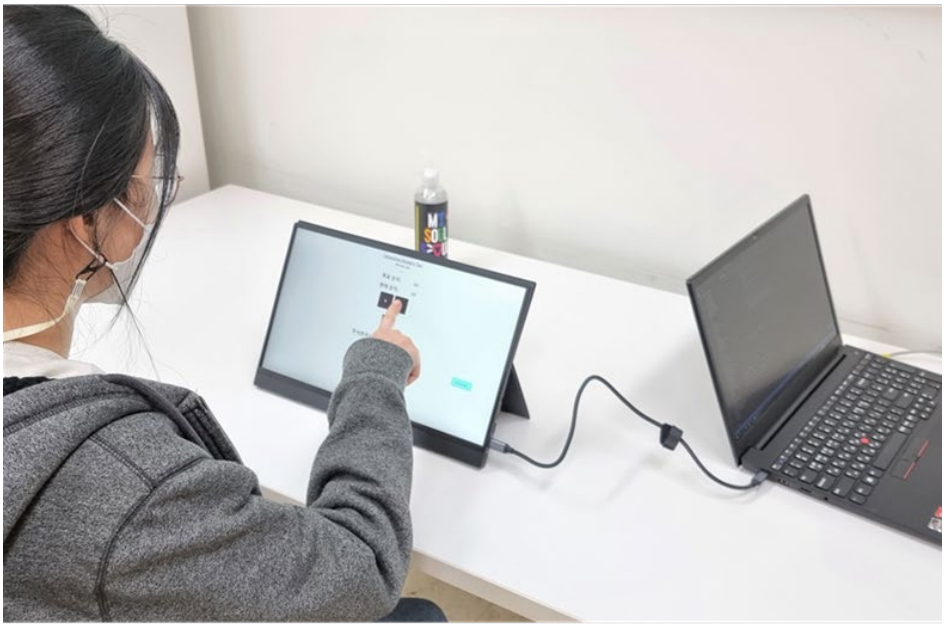
Thirty participants (17 males and 13 females) were recruited from the local university. All participants reporting an experience of VUI were Koreans and were between the ages of 20 and 38 (mean age = 27.6 years, SD = 4.47 years). One case was dropped due to failure to comply with the protocol and two cases were missing because of technical issues. All participants reported having a VUI experience and the usual VUI usage was balanced (daily: 6, weekly: 6, monthly: 8, rarely: 7) The experimental procedures, which were reviewed and approved by the university's Institutional Review Board (IRB No. 7001988-202210-HR-1709-02) (Appendix 4), were explained to each participant before beginning the experiment. All participants provided informed consent and received about 15$ for participation. Appendix 3.1 shows the questionnaire for the demographic information, and Appendix 1 shows the information of participants included the study 1.

5.6. Apparatus and Settings

The multimodal system program ran on a laptop, and the participants manipulated the program through a connected portable monitor which is a 15.6-in. touchscreen with a 1920 x 1080 resolution from ZEUSLAP. USB-C cable was used for the connection between the laptop and the monitor. The laptop was always connected to the Internet with LAN cable to use the speech Recognition API from google in web browser. Figure 5.10 shows the experimental equipment and the example scene. The participants were able to adjust the position of the monitor in order to touch it comfortably.



(a)



(b)

Figure 5.10. (a) Experimental setting of study 1 and (b) example scene of experiment.

5.7. Procedure

Overall procedure of study 1 is displayed in Figure 5.11. Before the experiment began, participants received a description of the experiment and a consent form. After that, they completed a demographic survey including questions about their age, gender, and usual frequency of VUI usage (daily, weekly, monthly, and rarely). The researcher explained specific instructions about each task. The order of the menu structures was non-hierarchy first followed by hierarchy, and the number of touches was presented randomly from 1 to 5 touches. Depending on the menu structure conditions, 5 or 8 levels were performed sequentially from high to low S/T, and each level consisted of 6 trials including one voice training trial. If participants performed all trials with the voice modality at a certain level, modality switching was considered to have occurred, and all subsequent levels had been replaced with voice.

After participants completed each level, SMEQ-based questionnaires were displayed to evaluate the physical and mental effort for operating each modality (Sauro & Dumas, 2009). Both efforts of voice modality were measured at all levels, but those efforts of touch modality were measured at random three times in non-hierarchy and four times in hierarchy. Then, the average of the three or four efforts data was calculated and used as representative values of efforts of touch modality. This was because the touch modal has a constant number of touches regardless of the level, so theoretically the efforts required should have a single value regardless of level.

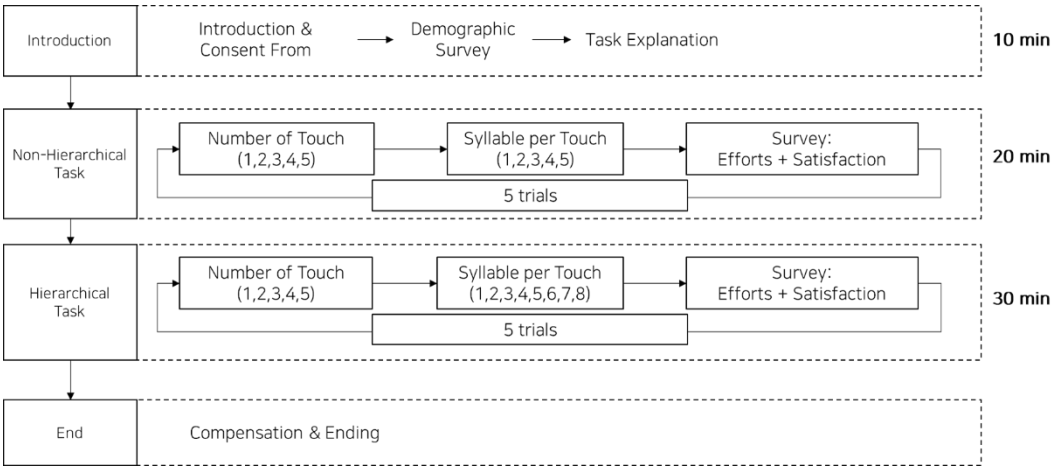


Figure 5.11. Overall procedure of study 1

5.8. Data Collection and Analysis

There are three dependent variables: one objective variable and two subjective variables. The modality that participants used to perform the tasks was recorded and used as the objective variable. Subjective variables were physical effort and mental effort, and they were items that modified SMEQ into physical and mental questionnaires for this experiment. Participants were asked to evaluate the use of the physical resource (moving hand, making sounds, etc.) and mental resources that are necessary to complete the trial with a specific modality. Both efforts were evaluated separately for each modality, and each questionnaire was a scale with 16 points between 0 to 150.

For the analysis of the used modality, five trials excluding the first voice training trial at each level were used, and up to 325 data lines were recorded for each participant. Binomial logistic regression analysis was conducted to determine whether the number of touches,

syllables per touch, and menu structures affect the usage of each modality.

ANOVA was performed to analyze the questionnaire data of the physical and mental effort required to operate each modality. When the significant effects of IVs were found, pairwise comparisons were conducted using the Games-Howell post-hoc test.

The significance level for all statistical tests was 0.05. All the statistical analyses were performed by using IBM SPSS Statistics version 26.0.

6. RESULTS

Data from 27 people (14 males and 13 females), excluding 3 people who had problems with data collection, were used for analysis, and 4 data lines, which caused errors (the task was not completed), were removed. The data for the non-hierarchy condition was 3347 lines, and the hierarchical condition was 5173 lines.

6.1. Modality Usage

Table 6.1 shows the descriptive statistics on the ratio of modality used for each experimental condition. It was found that the usage of voice increased as S/T decreased under both menu structures. In the result under each condition, the ‘modality switching’, which is the point at which the usage of voice and touch intersects, could be found, and it means that the S/T variable was appropriately controlled in this research.

Table 6.1. Descriptive statistics of voice and touch usage

Menu Structure	Number of Touches	Modality	Syllable per Touch (S/T)								Total
			8	7	6	5	4	3	2	1	
Non-Hierarchy	1	Touch				88 (65.2%)	74 (55.2%)	68 (50.4%)	62 (45.9%)	42 (31.1%)	334 (49.6%)
		Voice				47 (34.8%)	60 (44.8%)	67 (49.6%)	73 (54.1%)	93 (68.9%)	340 (50.4%)
	2	Touch				98 (73.1%)	90 (66.7%)	72 (53.3%)	55 (41.0%)	46 (34.1%)	361 (53.6%)
		Voice				36 (26.9%)	45 (33.3%)	63 (46.7%)	79 (59.0%)	89 (65.9%)	312 (46.4%)
	3	Touch				115 (85.2%)	100 (74.1%)	83 (61.5%)	59 (43.7%)	27 (20.0%)	384 (56.9%)
		Voice				20 (14.8%)	35 (25.9%)	52 (38.5%)	76 (56.3%)	108 (80.0%)	291 (43.1%)
	4	Touch				112 (83.0%)	94 (69.6%)	82 (60.7%)	48 (35.6%)	12 (8.9%)	348 (51.6%)
		Voice				23 (17.0%)	41 (30.4%)	53 (39.3%)	87 (64.4%)	123 (91.1%)	327 (48.4%)
	5	Touch				106 (81.5%)	92 (70.8%)	77 (59.2%)	45 (34.6%)	8 (6.2%)	328 (50.5%)
		Voice				24 (18.5%)	38 (29.2%)	53 (40.8%)	85 (65.4%)	122 (93.8%)	322 (49.5%)
Hierarchy	1	Touch	99 (73.3%)	96 (71.1%)	93 (68.9%)	89 (65.9%)	66 (48.9%)	51 (37.8%)	37 (27.4%)	34 (25.2%)	565 (52.3%)
		Voice	36 (26.7%)	39 (28.9%)	42 (31.1%)	46 (34.1%)	69 (51.1%)	84 (62.2%)	98 (72.6%)	101 (74.8%)	515 (47.7%)
	2	Touch	105 (77.8%)	94 (69.6%)	90 (66.7%)	59 (43.7%)	50 (37.0%)	36 (26.7%)	16 (11.9%)	8 (5.9%)	458 (42.4%)
		Voice	30 (22.2%)	41 (30.4%)	45 (33.3%)	76 (56.3%)	85 (63.0%)	99 (73.3%)	118 (88.1%)	127 (94.1%)	621 (57.6%)
	3	Touch	122 (90.4%)	112 (83.0%)	99 (73.3%)	72 (53.3%)	61 (45.2%)	27 (20.0%)	9 (6.7%)	7 (5.2%)	509 (47.1%)
		Voice	13 (9.6%)	23 (17.0%)	36 (26.7%)	63 (46.7%)	74 (54.8%)	108 (80.0%)	126 (93.3%)	128 (94.8%)	571 (52.9%)
	4	Touch	108 (83.1%)	103 (79.2%)	91 (70.0%)	66 (50.8%)	55 (42.3%)	31 (23.8%)	14 (10.8%)	5 (3.8%)	473 (45.5%)
		Voice	22 (16.9%)	27 (20.8%)	39 (30.0%)	64 (49.2%)	75 (57.7%)	99 (76.2%)	116 (89.2%)	125 (96.2%)	567 (54.5%)
	5	Touch	103 (90.4%)	83 (72.2%)	72 (62.6%)	62 (56.4%)	60 (54.5%)	36 (32.7%)	15 (13.6%)	5 (4.5%)	436 (48.8%)
		Voice	11 (9.6%)	32 (27.8%)	43 (37.4%)	48 (43.6%)	50 (45.5%)	74 (67.3%)	95 (86.4%)	105 (95.5%)	458 (51.2%)

The binomial logistic regression with forward selection (LR) was conducted to analyze the relationship between independent variables (the number of touches, syllables per touch, menu structure, VUI usage) and voice usage in the experiment. The menu structure and VUI usage were used to analyze as categorical variables. The results are summarized in Table 6.2. It was found that the number of touches did not significantly predict voice modality usage ($p = 0.115$). Syllable per touch(S/T) was a significant predictor, and, holding all other predictor variables constant, each additional increase of 1 S/T was associated with a 46.5% decrease in the odds of voice usage ($p < 0.001$, OR: 0.535, 95% CI: 0.519 – 0.551). It was also found that, holding all other predictor variables constant, the odds of voice usage in hierarchy increased by 3.451 times compared to non-hierarchy ($p < 0.001$, 95% CI: 3.078 – 3.869). In VUI usage, holding all other predictor variables constant, the odds of using voice modality increased by 308.5% ($p < 0.001$, 95% CI: 3.504 – 4.762) and 46.5% ($p < 0.001$, 95% CI: 1.269 – 1.691) for users who used VUI daily and weekly compared to using rarely. However, monthly VUI users used voice 36.8% less than rarely ($p < 0.001$, OR: 0.632, 95% CI: 0.552 – 0.723).

Table 6.2. Summary of logistic regression model of the modality usage.

IV		B	SE	Wald	*p*	OR	95% CI	
							Lower	Upper
Syllable per Touch (S/T)		-0.626	0.015	1648.604	.000***	0.535	0.519	0.551
Menu Structure[a]	**Hierarchy**	1.239	0.058	451.202	.000***	3.451	3.078	3.869
VUI Usage[b]	**Daily**	1.407	0.078	323.377	.000***	4.085	3.504	4.762
	Weekly	0.382	0.073	27.265	.000***	1.465	1.269	1.691
	Monthly	-0.459	0.069	44.48	.000***	0.632	0.552	0.723
Number of Touches		-	-	-	.115	-	-	-

***p<.001
reference group: [a]Non-hierarchy, [b]Rarely

The predicted voice usage of the logistic regression model was shown in Figure 6.1. 'Modality switching point', which is the point where the usage of voice and touch is reversed, was marked through the 50% auxiliary line in purple. It could be estimated that modality switching occurs at 2~3 S/T in non-hierarchy and at 4~5 S/T in hierarchy. The percentage correction between the model's prediction and the collected data was 74.0%. The Area Under the Curve (AUC) of the model derived from the ROC curve was 0.814, which means excellent discriminating ability (see Figure. 6.2)

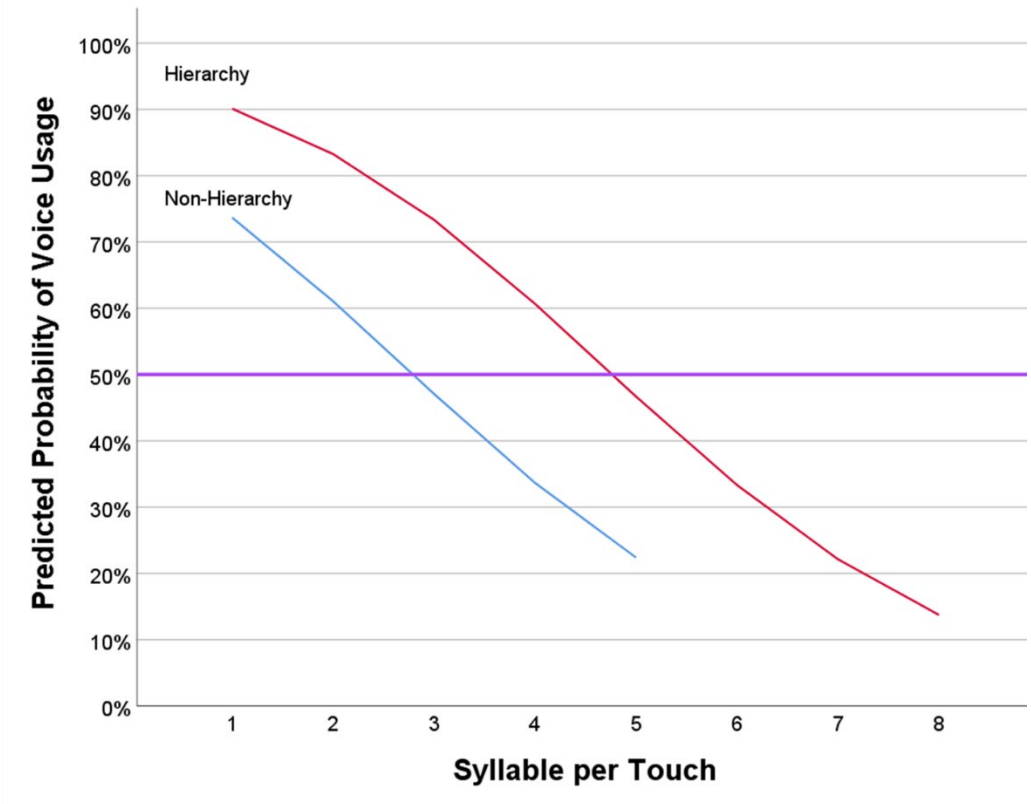


Figure 6.1. Probability of voice usage predicted by the logistic regression model

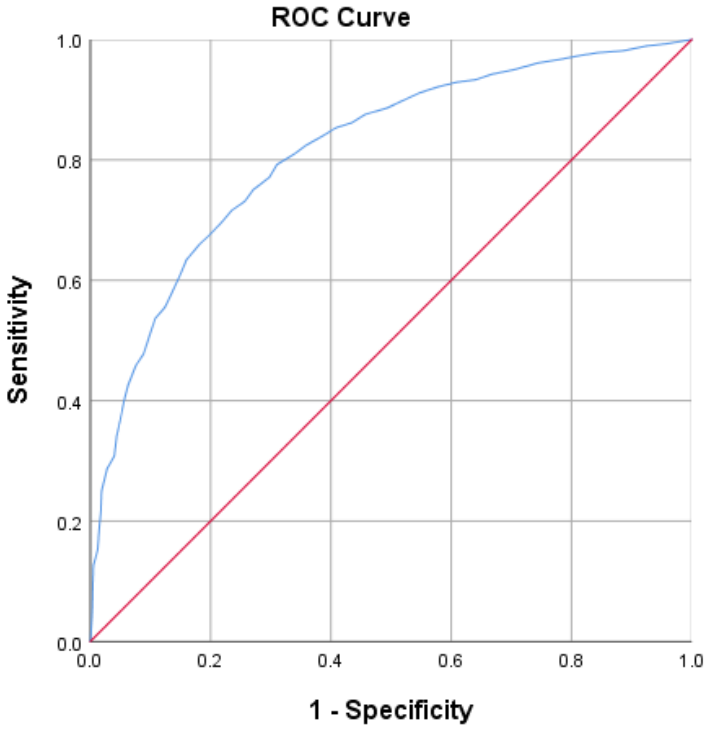


Figure 6.2. ROC curve of the logistic regression model of study 1

6.2. Interaction Efforts and Satisfaction

6.2.1. Interaction efforts and satisfaction of touch

Since touch modality had different types of tasks depending on the menu structure, changes in interaction efforts and satisfaction were explored for each menu structure.

Non-hierarchical Touch: Table 6.3, 6.4, and 6.5 show the descriptive statistics of physical effort, mental effort, and satisfaction of touch modality in non-hierarchy condition. Three measures were normally distributed, One-way ANOVA was conducted to determine the effect of the number of touches on physical effort, mental effort, and satisfaction.

First, in non-hierarchy condition, ANOVA was performed to compare the effect of the number of touches on both efforts of touch modality. The result revealed that there was a significant difference in physical effort between the number of touches ($F(4, 129) = 2.464$, $p = 0.048$). However, the pairwise comparison did not show the significance difference between number of touches.

There were no significant difference in mental effort and satisfaction between the number of touches [mental effort: $F(4, 129) = 1.538$, $p = 0.195$; satisfaction: $F(4, 129) = 0.647$, $p = 0.630$]. Therefore, in non-hierarchy condition, it couldn't be said that the mental effort touch task increased by the number of touches (see Figure 6.3).

Table 6.3. Result of ANOVA of physical effort of non-hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	15.9	10.33			
2	17.2	11.38			
3	20.5	11.82	4	2.464	0.048*
4	23.1	14.57			
5	25.7	17.82			
total	20.4	13.71			

*$p < 0.05$

Table 6.4. Result of ANOVA of mental effort of non-hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	19.6	16.29			
2	18.8	13.15			
3	24.8	15.15	4	1.538	0.195
4	26.0	17.02			
5	26.9	16.00			
total	23.2	15.70			

Table 6.5. Result of ANOVA of satisfaction of non-hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	70.4	19.81			
2	70.2	19.88			
3	66.9	21.15	4	0.647	0.630
4	64.3	20.81			
5	63.7	20.15			
total	67.1	20.26			

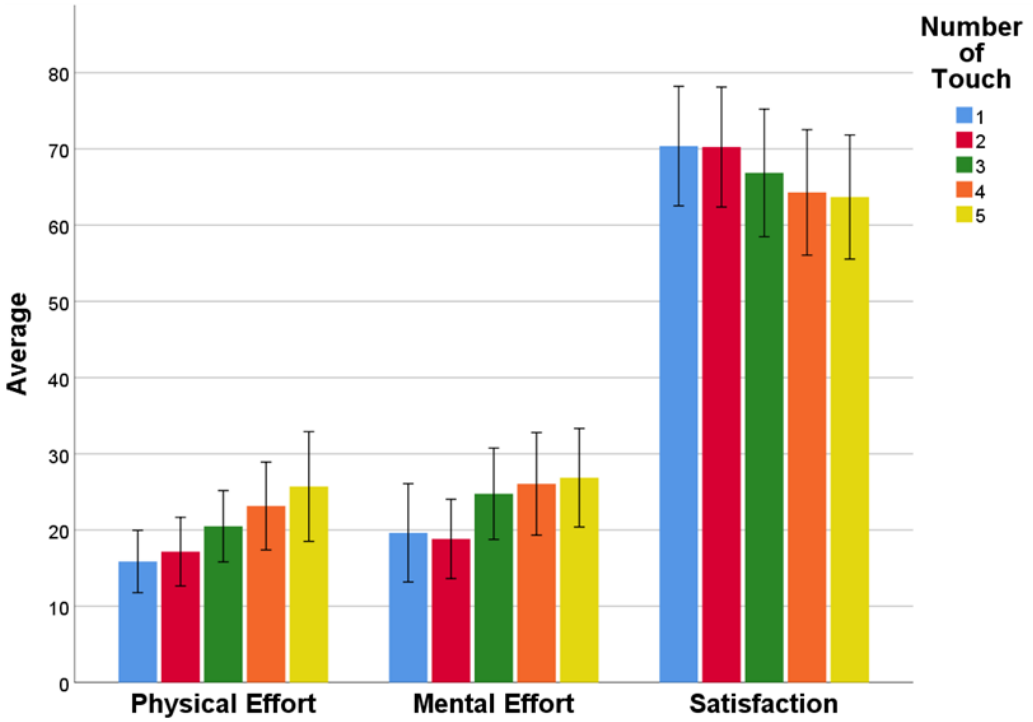


Figure 6.3. Interaction efforts and satisfaction of touch modality depending on number of touches in non-hierarchy condition

Hierarchical Touch Table 6.6, 6.7, and 6.8 show the descriptive statistics of physical effort, mental effort, and satisfaction of touch modality in hierarchy condition. In hierarchy condition, Welch's ANOVA was performed to compare the effect of the number of touches on both efforts of touch modality. The result revealed that there were significant differences in physical and mental efforts between the number of touches (Physical: $F(4, 59.699) = 8.201$, $p < 0.001$; Mental: $F(4, 60.837) = 5.049$, $p = 0.001$). However, there was no significant different in satisfaction ($F(4, 125) = 1.931$, $p = 0.109$). Figure 6.4 showed that the physical and mental efforts were increased by the number of touches.

Table 6.6. Result of ANOVA of physical effort of hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	19.6	11.26	4	8.201	0.000***
2	26.5	13.82			
3	31.0	17.51			
4	44.4	25.72			
5	44.5	27.81			
total	32.8	21.97			

***$p < 0.001$

Table 6.7. Result of ANOVA of mental effort of hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	25.9	16.53	4	5.049	0.001**
2	29.5	15.96			
3	35.2	19.55			
4	46.2	24.94			
5	48.1	27.37			
total	36.6	22.56			

**$p < 0.05$

Table 6.8. Result of ANOVA of satisfaction of hierarchy touch

Number of touches	M	SD	*df*	*F*	*p*
1	64.9	20.61	4	1.931	0.109
2	65.4	20.16			
3	61.3	19.62			
4	56.1	23.36			
5	51.2	24.83			
total	60.1	22.00			

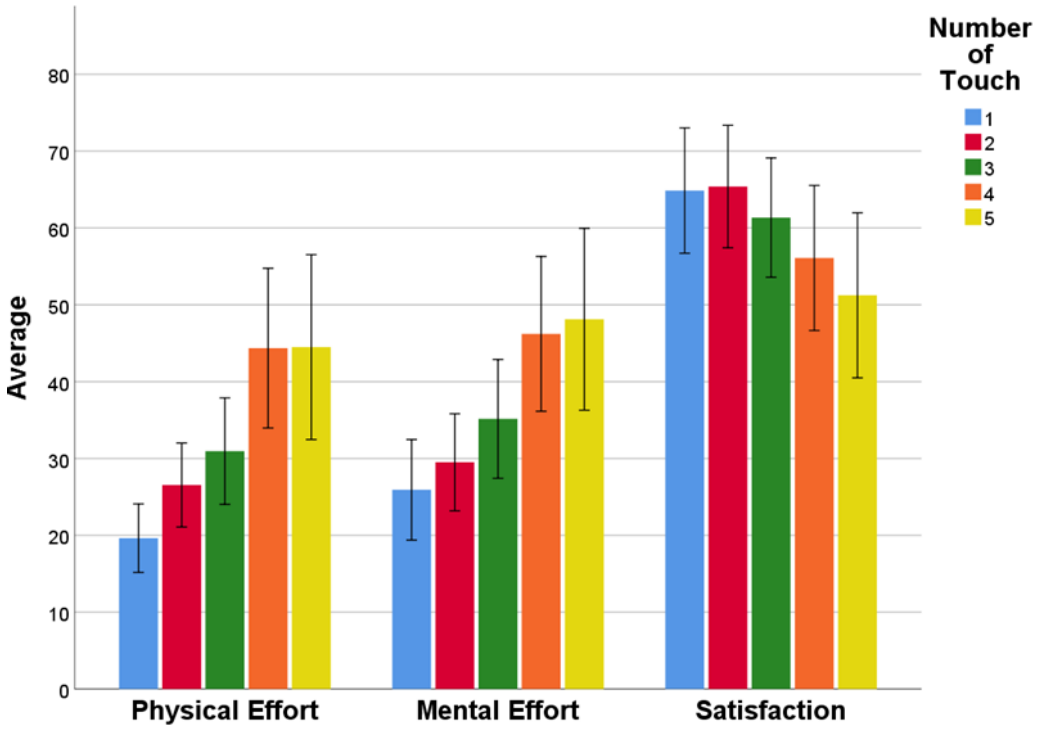


Figure 6.4. Interaction efforts and satisfaction of touch modality depending on the number of touches in hierarchy condition

6.2.2. Interaction efforts and satisfaction of voice

Physical Effort: The three-way ANOVA was conducted to determine the effects of number of touches, syllable per touch and menu structure on physical effort of voice modality. Table 6.9 shows the summary of ANOVA.

The result of ANOVA with physical effort showed that the main effects of number of touches and syllable per touch were significant [number of touches: $F(4, 1340) = 96.699$, $p < 0.001$; syllable per touch: $F(7, 1340) = 35.759$, $p < 0.001$; menu structure: $F(1, 1340) = 0.505$, $p = 0.478$], and only the number of touches × syllable per touch effect was significant among the interaction effects [number of touches × syllable per touch: $F(28, 1340) = 2.244$, $p < 0.001$; syllable per touch × menu structure: $F(4, 1340) = 0.337$, $p = 0.853$; number of touches × menu structure: $F(4, 1340) = 0.113$, $p = 0.978$; number of touches × syllable per touch × menu structure: $F(16, 1340) = 0.172$, $p = 1.000$].

Table 6.9. Result of three-way ANOVA of Physical Effort of voice modality

DV	Variables	*df*	*F*	*p*	η^2
Physical Effort	Number of touches	4	96.699	0.000***	0.224
	Syllable per touch	7	35.759	0.000***	0.157
	Menu structure	1	0.505	0.478	0.000
	Number of touches × syllable per touch	28	2.244	0.000***	0.045
	Syllable per touch × menu structure	4	0.337	0.853	0.001
	Number of touches × menu structure	4	0.113	0.978	0.000
	Number of touches × syllable per touch × menu structure	16	0.172	1.000	0.002

***$p < 0.001$

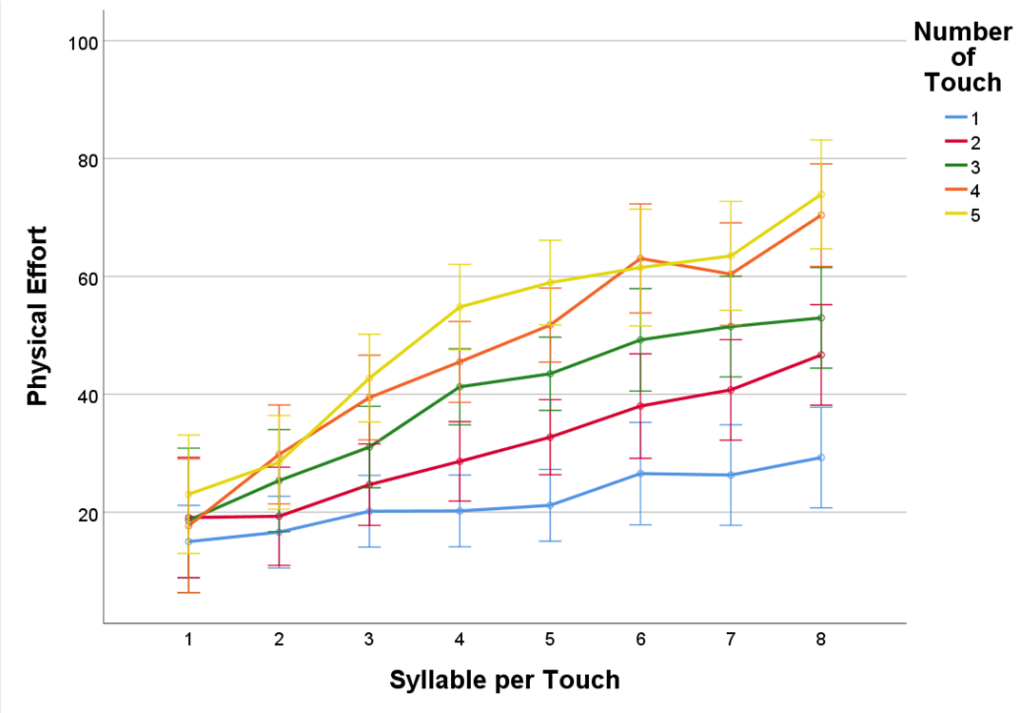


Figure 6.5. Physical effort of voice modality according to number of touches and syllables per touch

Mental Effort: The three-way ANOVA was conducted to determine the effects of number of touches, syllable per touch and menu structure on mental efforts of voice modality. Table 6.10 shows the summary of ANOVA.

The result of ANOVA with mental effort showed that the main effects of number of touches and syllable per touch were significant [number of touches: $F(4, 1340) = 67.453$, $p = 0.000$; syllable per touch: $F(7, 1340) = 29.317$, $p = 0.000$; menu structure: $F(1, 1340) = 0.163$, $p = 0.687$], and only the number of touches × syllable per touch effect was

significant among the interaction effects [number of touches × syllable per touch: F(28, 1340) = 1.815, p = 0.006; syllable per touch × menu structure: F(4, 1340) = 0.572, p = 0.683; number of touches × menu structure: F(4, 1340) = 0.171, p = 0.953; number of touches × syllable per touch × menu structure: F(16, 1340) = 0.134, p = 1.000].

Table 6.10. Result of three-way ANOVA of Mental Effort of voice modality

DV	Variables	*df*	*F*	*p*	η^2
Mental Effort	Number of touches	4	67.453	0.000***	0.168
	Syllable per touch	7	29.317	0.000***	0.133
	Menu structure	1	0.163	0.687	0.000
	Number of touches × syllable per touch	28	1.815	0.006***	0.037
	Syllable per touch × menu structure	4	0.572	0.683	0.002
	Number of touches × menu structure	4	0.171	0.953	0.001
	Number of touches × syllable per touch × menu structure	16	0.134	1.000	0.002

***$p < 0.001$

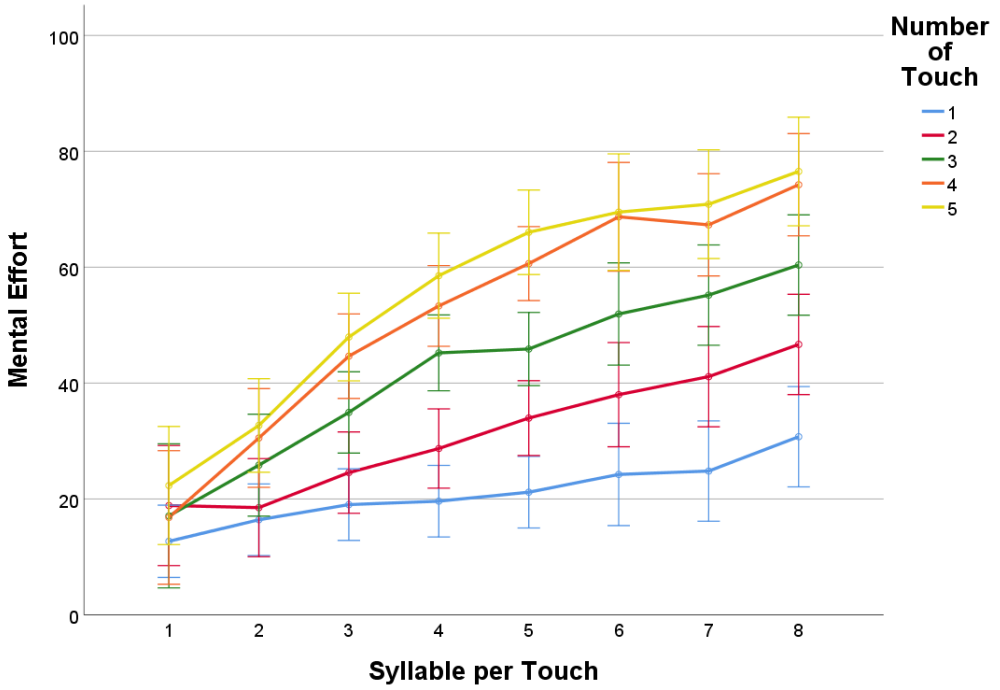


Figure 6.6. Mental effort of voice modality according to number of touches and syllables per touch

Satisfaction: The three-way ANOVA was conducted to determine the effects of number of touches, syllable per touch and menu structure on satisfaction of voice modality. Table 6.11 shows the summary of ANOVA.

The result of ANOVA with satisfaction showed that the main effects of number of touches and syllable per touch were significant [number of touches: $F(4, 1340) = 24.069$, $p = 0.000$; syllable per touch: $F(7, 1340) = 12.196$, $p = 0.000$; menu structure: $F(1, 1340) = 0.023$, $p = 0.879$], and the interaction effects were not significant [number of touches

× syllable per touch: F(28, 1340) = 0.502, p = 0.986; syllable per touch × menu structure: F(4, 1340) = 0.105, p = 0.981; number of touches × menu structure: F(4, 1340) = 0.609, p = 0.656; number of touches × syllable per touch × menu structure: F(16, 1340) = 0.142, p = 1.000].

Table 6.11. Result of three-way ANOVA of Satisfaction of voice modality

DV	Variables	*df*	*F*	*p*	η^2
Satis-faction	Number of touches	4	24.069	0.000***	0.067
	Syllable per touch	7	12.196	0.000***	0.060
	Menu structure	1	0.023	0.879	0.000
	Number of touches × syllable per touch	28	0.502	0.986	0.010
	Syllable per touch × menu structure	4	0.105	0.981	0.000
	Number of touches × menu structure	4	0.609	0.656	0.002
	Number of touches × syllable per touch × menu structure	16	0.142	1.000	0.002

***p < 0.001

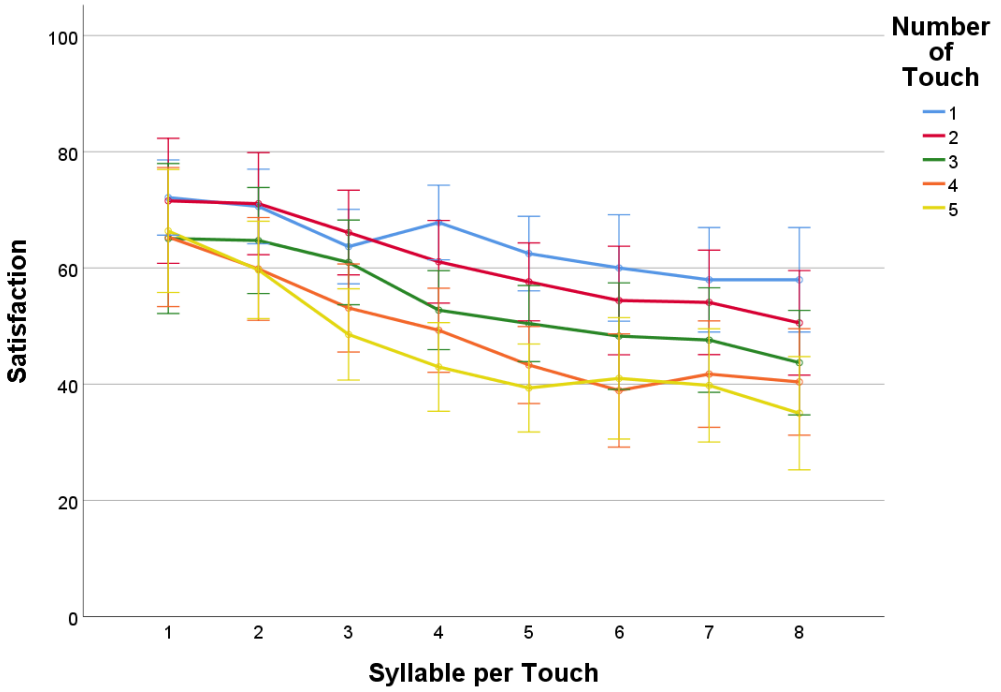


Figure 6.7. Satisfaction of voice modality according to number of touches and syllables per touch

6.2.3. Interaction efforts and satisfaction between menu structure and modality

Two-ANOVA was performed to determine the difference in the effects of menu structure on interaction efforts and satisfaction of each modality. If there were an interaction between menu structure and modality, a simple main effect analysis was conducted to analyze the effects of menu structure within voice or touch.

Physical Effort The result of ANOVA with physical effort showed that the modality and menu structure effects were significant [modality: $F(1, 1285) = 8.889$, $p = 0.003$; menu structure: $F(1, 1285) = 14.038$, $p = 0.000$], and the interaction effect was no significant [modality × menu structure: $F(1, 1285) = 15.465$, $p = 0.000$] (Table 6.12). Only the simple main effect of menu structure on touch was significant [in voice: $F(1, 1285) = 0.041$, $p = 0.840$; in touch: $F(1, 1285) = 18.694$, $p = 0.000$] (Table 6.13)

Table 6.12. Result of two-way ANOVA for physical effort by modality and menu structure

DV	Variables	*df*	*F*	*p*	η^2
Physical Effort	Modality	1	8.998	0.003**	0.007
	Menu structure	1	14.038	0.000***	0.011
	Modality × Menu structure	1	15.465	0.000***	0.012

**$p < 0.01$, ***$p < 0.001$

Table 6.13. Simple main effect of menu structure on physical effort in both modalities

Modality	Non-hierarchy (SD)	Hierarchy (SD)	*df*	*F*	*p*	η^2
Voice	31.6 (0.94)	31.3 (1.14)	1	0.041	0.840	0.000
Touch	20.4 (2.00)	32.8 (2.03)	1	18.694	0.000***	0.014

***$p < 0.001$

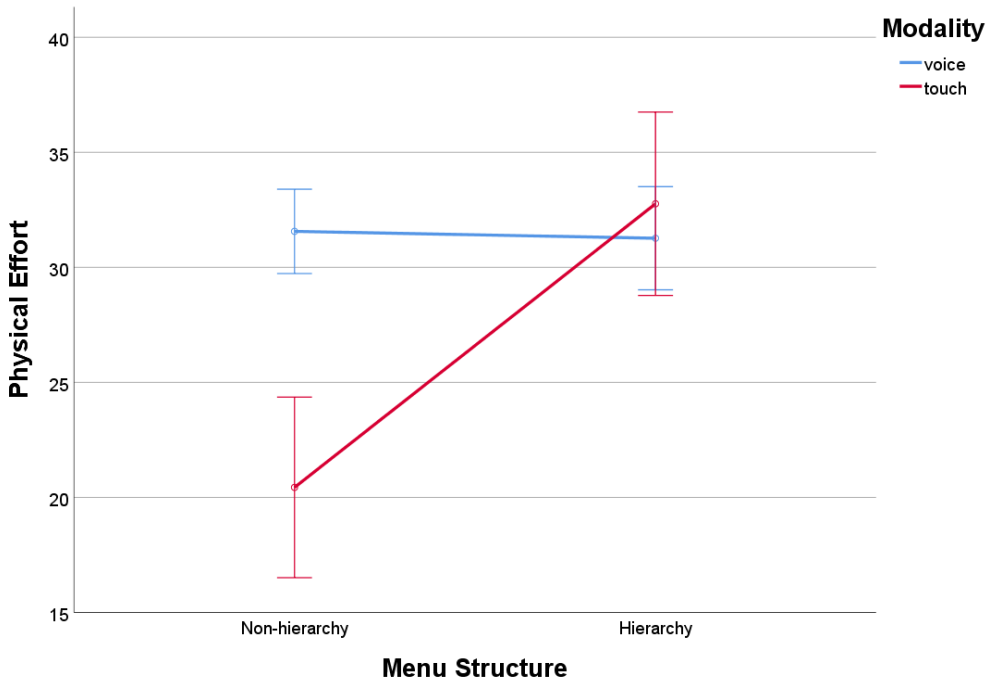


Figure 6.8. Difference in physical effort by modality according to menu structure

Mental Effort The result of ANOVA with mental effort showed that the modality and menu structure effects were significant [modality: F(1, 1285) = 4.080, p = 0.044; menu structure: F(1, 1285) = 12.975, p = 0.000], and the interaction effect was no significant [modality × menu structure: F(1, 1285) = 16.905, p = 0.000] (Table 6.14). Only the simple main effect of menu structure on touch was significant [in voice: F(1, 1285) = 0.307, p = 0.580; in touch: F(1, 1285) = 18.862, p = 0.000] (Table 6.15).

Table 6.14. Result of two-way ANOVA for mental effort by modality and menu structure

DV	Variables	*df*	*F*	*p*	η^2
Mental Effort	Modality	1	4.080	0.044*	0.003
	Menu structure	1	12.975	0.000***	0.010
	Modality × menu structure	1	16.905	0.000***	0.013

*$p < 0.05$, ***$p < 0.001$

Table 6.15. Simple main effect of menu structure on mental effort in both modalities

Modality	Non-hierarchy (SD)	Hierarchy (SD)	*df*	*F*	*p*	η^2
Voice	33.8 (1.01)	32.9 (1.23)	1	0.307	0.580	0.000
Touch	23.2 (2.16)	36.6 (2.19)	1	18.862	0.000***	0.014

***$p < 0.001$

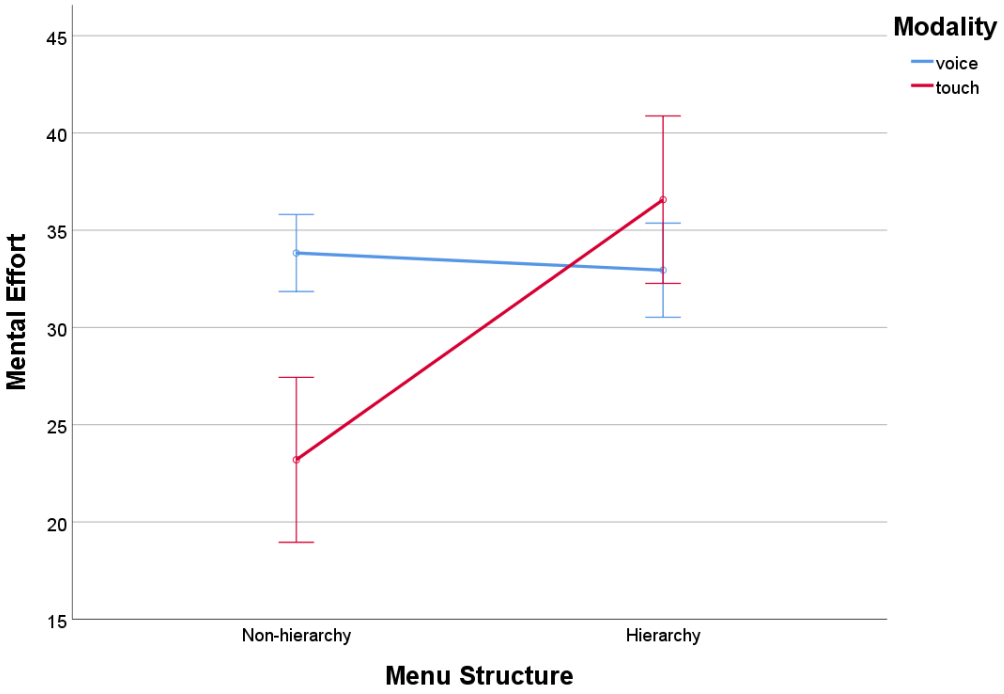


Figure 6.9. Difference in mental effort by modality according to menu structure

Satisfaction The result of ANOVA with satisfaction showed that the modality and menu structure effects were significant [modality: F(1, 1285) = 7.483, $p = 0.006$; menu structure: F(1, 1285) = 3.954, $p = 0.047$], and the interaction effect was no significant [modality × menu structure: F(1, 1285) = 4.981, $p = 0.026$] (Table 6.16). Only the simple main effect of menu structure on touch was significant [in voice: F(1, 1285) = 0.070, $p = 0.791$; in touch: F(1, 1285) = 5.646, $p = 0.018$] (Table 6.17).

Table 6.16. Result of two-way ANOVA for satisfaction by modality and menu structure

DV	Variables	*df*	*F*	*p*	η^2
Satis-faction	Modality	1	7.483	0.006**	0.006
	Menu structure	1	3.954	0.047*	0.003
	Modality × menu structure	1	4.981	0.026*	0.004

*$p < 0.05$, **$p < 0.01$

Table 6.17. Simple main effect of menu structure on satisfaction in both modalities

Modality	Non-hierarchy (SD)	Hierarchy (SD)	*df*	*F*	*p*	η^2
Voice	58.8 (0.97)	59.2 (1.19)	1	.070	.791	0.000
Touch	67.1 (2.08)	60.1 (2.11)	1	5.646	.018*	0.004

*$p < 0.05$

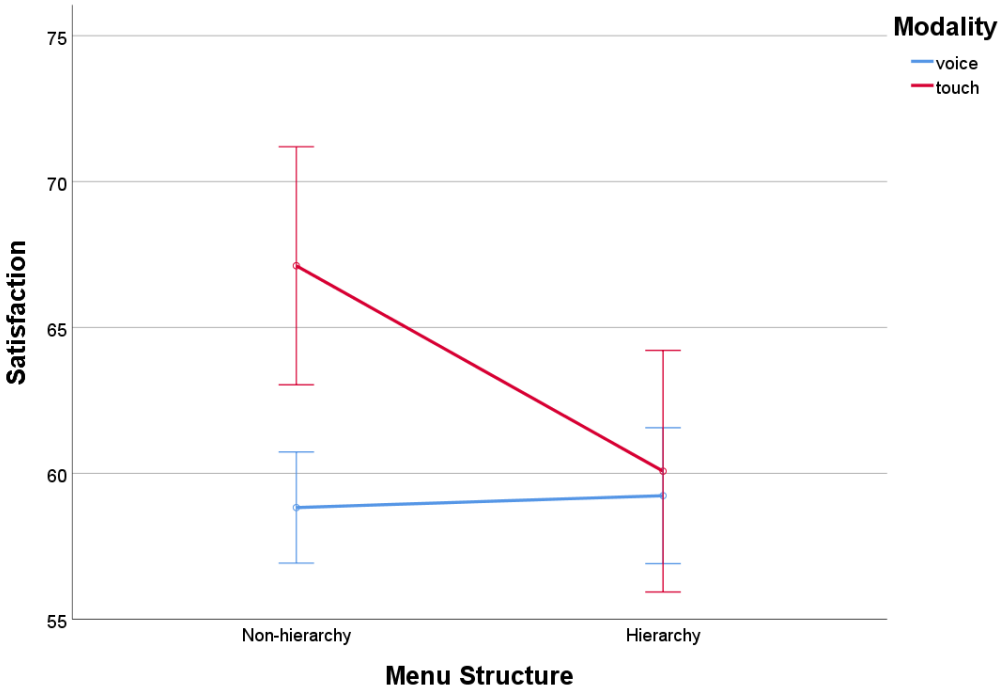


Figure 6.10. Difference in satisfaction by modality according to menu structure

6.2.4. Touch modality efforts between menu structures

To analyze the cause of change in the 'modality switching point' according to the menu structure, T-test was performed using both efforts data of touch only (see Table 6.18). Both touch efforts showed a significant difference between menu structures ($p_{physical}$=.000, p_{mental}=.000). Therefore, physical and mental efforts of touch in hierarchy were significantly higher than in non-hierarchy.

Table 6.18 T-test results of physical effort and mental effort of touch modality between menu structures.

DV	Menu structure	N	Mean	SD	*t(p)*
Physical Effort	Non-Hierarchy	134	20.44	13.713	-5.576(.000)***
	Hierarchy	130	32.76	21.966	
Mental Effort	Non-Hierarchy	134	23.20	15.701	-5.451(.000)***
	Hierarchy	130	36.57	22.557	

***p<.001

6.3. Modality Switching Points

Welch's ANOVA was performed to verify the relationship between interaction efforts, satisfaction and modality switching point. Table 6.19 shows the results of ANOVA for modality on interaction efforts and satisfaction in both menu structures.

To compare the efforts of touch and voice, a modality which is a variable consisting of 'touch' and ' syllables per touch' was created (Modality: Touch, 1, 2, 3, 4, 5, 6, 7, 8). Welch's ANOVA was performed to compare the effect of modality on physical effort, mental effort, and satisfaction (see Table 6.19). In non-hierarchy condition, physical effort had a significant difference between modalities ($F(5, 337.242) = 19.921$, $p = 0.000$). There was a significant difference in mental effort between modalities ($F(5, 340.215) = 21.981$, $p = 0.000$). Satisfaction also showed a significant difference between modalities ($F(5, 348.762) = 12.294$, $p = 0.000$).

In hierarchy condition, there were significant differences in three dependent variables. Physical effort showed a significant difference between modalities ($F(8, 347.582) = 31.154$, $p = 0.000$). Mental effort also had a significant difference between modalities ($F(8, 349.146) = 39.888$, $p = 0.000$). And satisfaction had a significant difference $F(8, 340.546) = 13.527$, $p = 0.000$).

Games-Howell post-hoc for multiple comparison were conducted to find the difference in three dependent variables between modalities.

Table 6.19. results of ANOVA for modality on interaction efforts and satisfaction in both menu structures

Menu Structure	DV	Modality (Mean (SD))									*F*	*p*
		Touch	Voice									
			1 S/T	2 S/T	3 S/T	4 S/T	5 S/T	6 S/T	7 S/T	8 S/T		
Non-hierarchy	Physical Effort	20.4 (13.71)	19.6 (18.22)	23.7 (16.45)	32.9 (23.72)	36.4 (26.52)	41.3 (30.80)	-	-	-	19.921	.000***
	Mental Effort	23.2 (15.70)	18.5 (16.95)	25.4 (20.29)	35.1 (24.59)	40.2 (29.97)	45.1 (34.04)	-	-	-	21.981	.000***
	Satisfaction	66.6 (21.75)	69 (21.81)	64.7 (23.26)	57.9 (24.94)	55.2 (25.52)	50.6 (27.69)	-	-	-	12.294	.000***
Hierarchy	Physical Effort	32.8 (21.97)	16.7 (12.77)	22.2 (14.44)	28.3 (18.54)	37.4 (25.18)	40.1 (24.92)	46.7 (26.16)	47.9 (29.09)	53.9 (32.96)	31.154	.000***
	Mental Effort	36.6 (22.56)	14.6 (12.37)	22.1 (14.72)	30.7 (19.41)	39.3 (23.86)	43.8 (26.02)	49.2 (27.94)	51.2 (30.64)	57 (33.73)	39.888	.000***
	Satisfaction	61.2 (23.26)	69.9 (19.14)	67 (18.61)	60.2 (22.73)	55.9 (23.83)	52 (24.91)	49.1 (25.02)	48.5 (25.93)	45.9 (25.73)	13.527	.000***

***p<.001

Physical effort showed a gradual increase as S/T increased in voice modality. The physical efforts of each modality in both menu structures were shown in Figure 6.11, and the summary of post-hoc analysis was shown in Table 6.20. In non-hierarchy, the physical effort of touch showed no significant difference with voice modality in 1 and 2 S/T, and was significantly lower than that of voice modality in 3, 4, and 5 S/T ($p_1 = 0.999$, $p_2 = 0.54$, $p_{3,4,5} < 0.005$). In hierarchy, Games-Howell post-hoc result revealed that physical effort of touch did not show a significant difference with voice modality in 3, 4, and 5 S/T ($p_3 = 0.795$, $p_4 = 0.874$, $p_5 = 0.292$). However, physical efforts of voice modality in 1 and 2 S/T were lower, and in 6, 7, and 8 S/T were higher in physical effort than that of touch ($p_{1,2,6,7,8} < 0.005$).

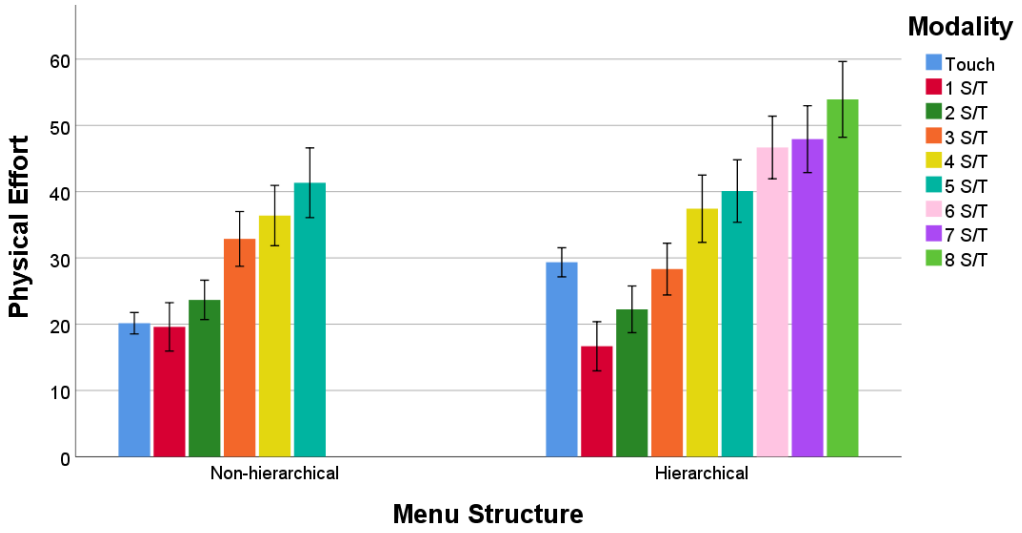


Figure 6.11. Physical effort by modality in each menu structure

Table 6.20. Result of Games-Howell post-hoc analysis on physical effort

	Games-Howell Post-hoc								
Non-hierarchy	1[a]	Touch[a]	2[a]	3[b]	4[b]	5[b]			
Hierarchy	1[a]	2[ab]	3[bc]	Touch[cd]	4[cde]	5[de]	6[ef]	7[ef]	8[f]

Mental effort also showed a gradual increase as S/T increased in voice modality. The mental efforts of each modality in both menu structures were shown in Figure 6.12, and the summary of post-hoc analysis was shown in Table 6.21. In non-hierarchy, the result of post-hoc revealed that the mental effort of touch did not significantly differ from voice modality in 1 and 2 S/T and was significantly lower than that of voice in 3, 4, and 5 S/T ($p_1 = 0.259$, $p_2 = 0.928$, $p_{3,4,5} < 0.005$). In hierarchy, mental effort of touch was not significantly different from voice in 3, 4, and 5 S/T, but there were significantly differences with voice in 1, 2, 6, 7, and 8 S/T ($p_3 = 0.500$, $p_4 = 0.994$, $p_5 = 0.356$, $p_{1,2,6,7,8} < 0.005$).

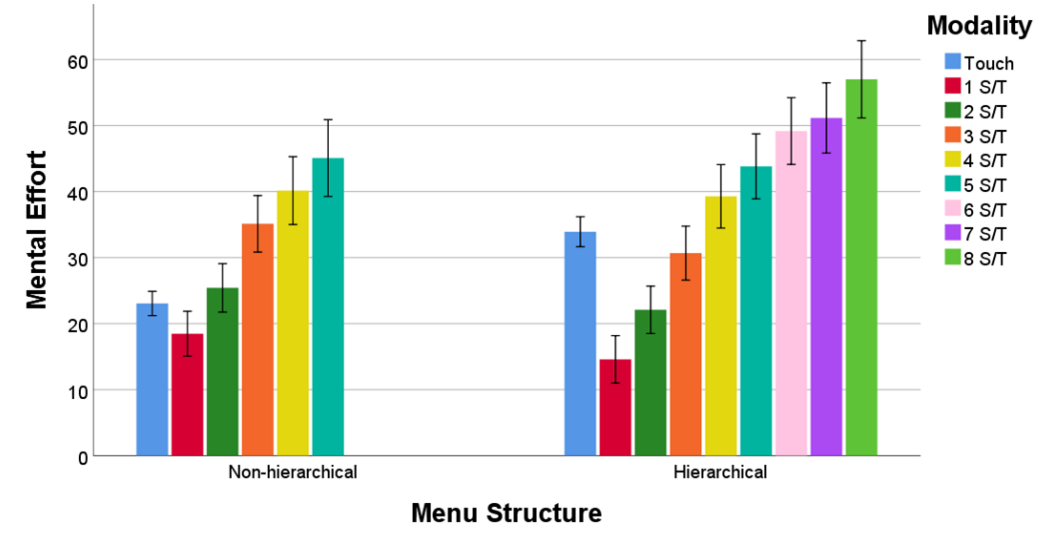


Figure 6.12. Mental effort by modality in each menu structure

Table 6.21. Result of Games-Howell post-hoc analysis on mental effort

Games-Howell Post-hoc									
Non-hierarchy	1^{a}	Touch^{a}	2^{a}	3^{b}	4^{b}	5^{b}			
Hierarchy	1^{a}	2^{ab}	3^{bc}	Touch^{cd}	4^{cdf}	5^{def}	6^{efg}	7^{fg}	8^{g}

Satisfaction gradually decreased as S/T increased in voice modality. The satisfactions of each modality in both menu structures were shown in Figure 6.13, and the summary of post-hoc analysis was shown in Table 6.22. In non-hierarchy, the satisfaction of touch did not significantly differ from that of voice modality in 1 and 2 S/T ($p_1 = 0.927$, $p_2 = 0.967$). And the satisfaction of touch was significantly higher than that of voice modality in 3, 4, and 5 S/T ($p_3 = 0.007$, $p_{4,5} < .005$). In hierarchy, the satisfaction of touch was not different from that of voice modality in 1, 2, 3, and 4 S/T ($p_1 = 0.107$, $p_2 = 0.365$, $p_3 = 1.000$, $p_4 = 0.587$) and was significantly higher than that of voice modality in 5, 6, 7, and 8 S/T ($p_5 = 0.019$, $p_{6,7,8} < .005$)

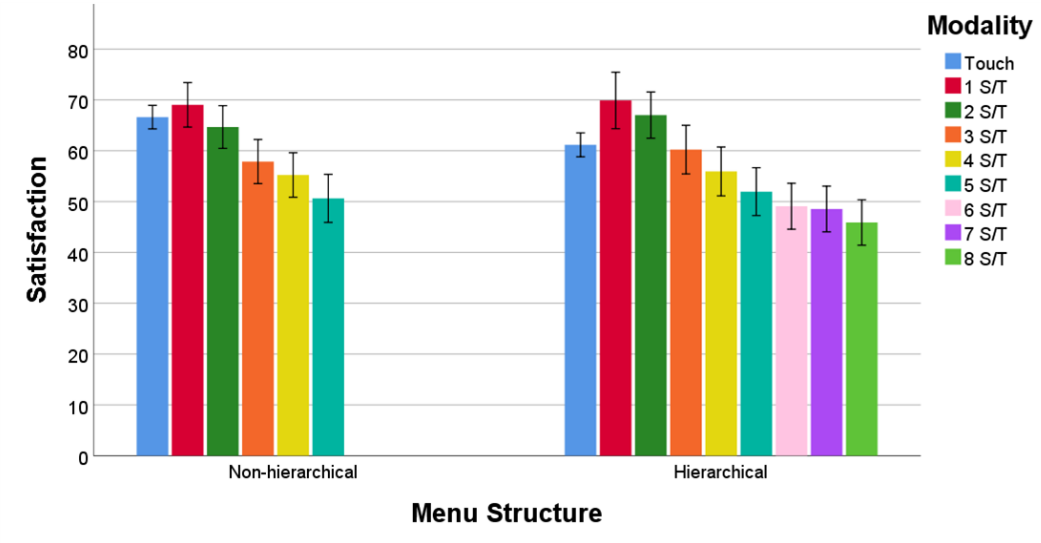


Figure 6.13. Satisfaction by modality in each menu structure

Table 6.22. Result of Games-Howell post-hoc analysis on satisfaction

	Games-Howell Post-hoc								
Non-hierarchy	1^{a}	Touch^{a}	2^{ab}	3^{bc}	4^{c}	5^{c}			
Hierarchy	1^{a}	2^{a}	Touch^{ab}	3^{abc}	4^{bcd}	5^{cd}	6^{d}	7^{d}	8^{d}

7. DISCUSSION

The results of this study revealed that users' modality usage varies by the system's menu structure and the features of modality in a multimodal system. This was significant in that the modality usage was not determined by the individual characteristics of voice and touch, but rather that there was a certain ratio according to the relationship between the two modalities. The fact that modality switching occurs in a higher number of S/T in the hierarchical task than in the non-hierarchical task proved that users perceived the utility of the voice and touch differently depending on the composition of multimodal systems or functions.

First of all, it was figured out that users' modality selection was determined by syllable per touch (S/T), the ratio of voice and touch unit, and that one touch could be replaced with several syllables. This supported the findings of some studies that voice had higher usability in text entry tasks than touch (A. L. Cox et al., 2008). In terms of language, a syllable consists of more than one letter and, basically, the input speed in speech is also much faster than in touch, so, in a way, it is natural that voice has higher usability in text entry. However, in actual systems, many functions require users to speak more syllables than the number of touches, that is, more than one syllable per touch. As in Figure 6.1, modality switching (intersection of voice and touch usage) occurred in non-hierarchy at 2~3 S/T and hierarchy at 4~5 S/T. This implies that such modality switching can occur not only in text entry but also in various tasks of the actual system.

On the other hand, it was an interesting result that there was no effect of the number of touches on modality usage. In previous studies, the number of touches was regarded as the interaction step of touch and was suggested to be an important factor (Schaffer et al., 2011), but this study showed slightly different results. The result revealed that modality usage was not simply affected by the modality features, but by certain ratios between them. This also seems to be a different result from studies that emphasized natural speech for users (Hua & Ng, 2010). To make users speak naturally, as in those studies, many products on the market emphasize that users can input commands in complete sentences. Even though it is a natural command, making the user speak longer to perform the same function may discourage the user from using voice. It is indisputably important to make them speak naturally within the same conditions. But managing the speech so that it does not cross the modality switching point may be more important to increase the user's voice usage.

The differences in the interaction efforts and satisfaction of both modalities could explain and support this modality usage. The increase in the number of touches had made it harder to use voice and touch modalities. As the touch modality became more difficult, the voice modality became equally difficult, so the number of touches would not have a significant effect on the modality selection. In terms of satisfaction, slightly different results were obtained. The satisfaction of touch modality was not affected by number of touches, whereas the satisfaction of voice modality decreased as number of touches increased. This was a different result from the interactive efforts. This means that, as revealed in the previous TAM study (Chao, 2019), satisfaction can act as a secondary factor rather than a

primary cognitive factor such as interaction effort. Therefore, a separate consideration of satisfaction is required in future studies.

The modality switching in voice and touch could also be proved indirectly in the analysis results of physical effort, mental effort, and satisfaction. In non-hierarchy and hierarchy, below the modality switching point, both efforts of voice modality were lower than or equal to those of touch modality, and above the point, both efforts of voice were higher than or equal to those of touch modality. On the other hand, the satisfaction of touch modality did not differ from that of voice modality below the modality switching point. And above the modality switching point, the satisfaction of touch was significantly higher than that of voice. This result was in line with the research that suggested that, as the costs of interaction were driven down, more users participated in the interaction (L. Hong et al., 2008). Although not perfectly clear, users were aware of the efforts required to operate each modality, which can be closely related to users' modality usage in multimodal systems. Therefore, users were able to evaluate the interaction efforts required for voice and touch through repeated operating experiences, and they selected the modality by comparing each interaction effort.

The user's usual VUI usage also affected the modality selection. As the results of this study, it was found that, in general, the more usual VUI usage was, the more voice modality was selected. People who usually use voice recognition a lot would have higher intimacy, proficiency, and trust in voice recognition, and accordingly, they may have used voice more

than touch (Rupp et al., 2018). However, monthly users had lower voice usage than rarely users, which may be influenced by factors other than the aforementioned factors. Therefore, further study is needed to identify which factors of attitude toward VUI are correlated with usual VUI usage and to investigate their effects of them on voice modality usage in multimodal systems.

CHAPTER IV: STUDY 2

8. RESEARCH MODEL

From the previous chapter, we learned that menu structure and modality features affect users' modality selection in multimodal systems. And we found how physical effort, mental effort, and satisfaction of the voice and touch are affected by menu structure and modality features. Based on the results of study 1, study 2 aimed to figure out the effects of various contexts on the user's modality selection, interaction efforts, and satisfaction. This study tried to verify that the interaction efforts and satisfaction of each modality would change in certain multitasking situations, and the user's modality selection also changes accordingly. The research model of the study 2 was presented in Figure 8.1.

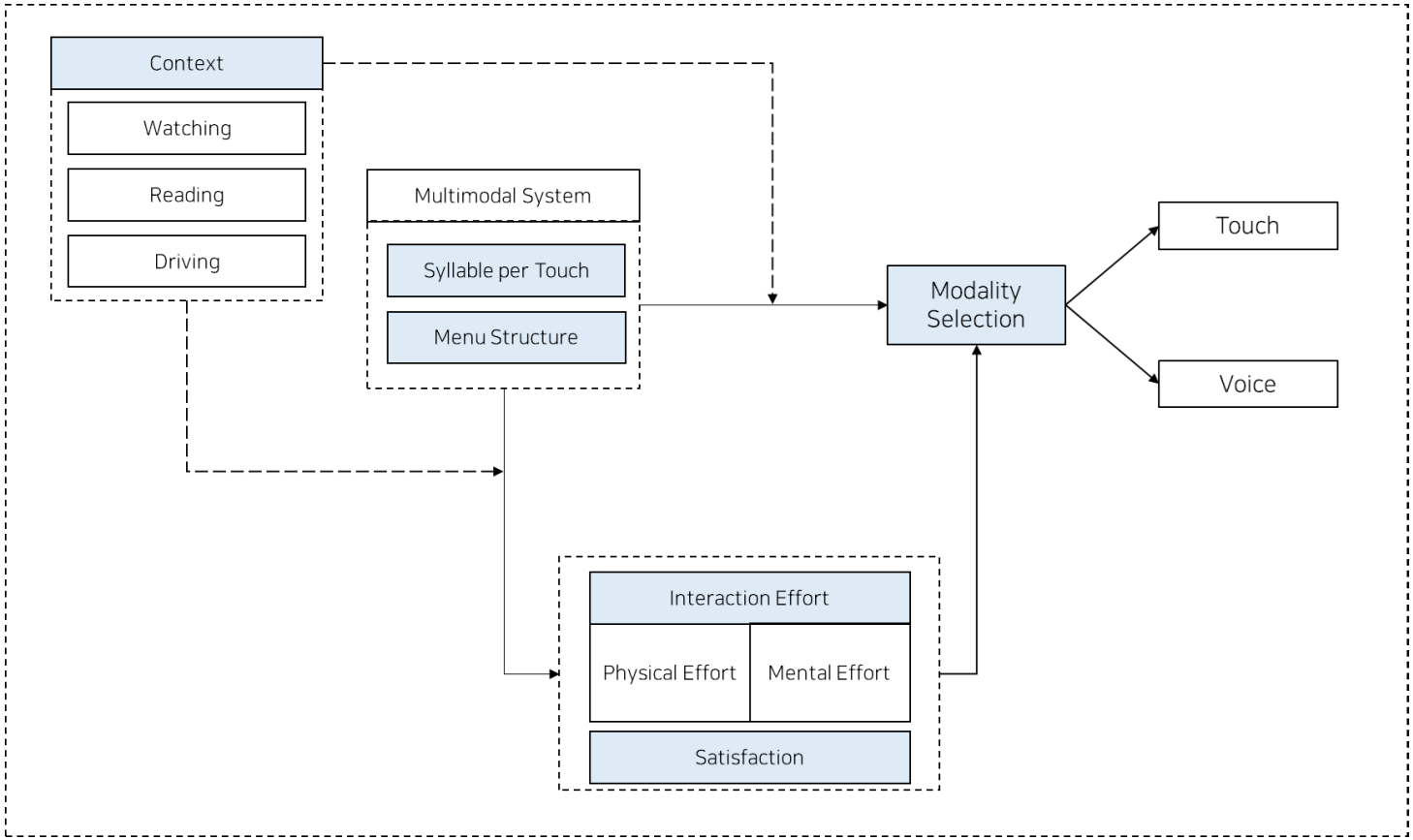


Figure 8.1. The research model of the study 2

9. OBJECTIVE AND HYPOTHESES

9.1. Objective

As mentioned in Section 8, the object of this study was to investigate the effects of contexts on the user's modality selection in multimodal systems, and the following research questions were presented below:

- RQ1: How will contexts affect the users' modality selection?
- RQ2: How will contexts affect interaction efforts and satisfaction?

To achieve the goal of this study, some hypotheses were developed and tested. An experiment with a structure similar to study 1 including various contexts was conducted to test the hypotheses.

9.2. Hypotheses

Hypothesis 5: The multitasking contexts would affect the user's modality selection.

H5a: The usage of voice modality would be higher in multitasking contexts (watching, reading, and driving) than in baseline.

H5b: The effect of physical resources on the voice usage would be stronger than that of mental resources.

H5c: The changes in the voice usage by multitasking contexts would be different by the menu structure.

There are many studies which showed that voice modality has high usability in various contexts (Beckers et al., 2014; S. Cox et al., 2003; He et al., 2015; Luria et al., 2017). Multitasking situations limit the resources available to users. Accordingly, in different contexts of use, it was expected that people would use voice modality more than touch modality because that requires less physical and mental effort to manipulate.

According to the multiple resource model by Wickens (2002), overlapping of the same resource causes users to process information sequentially, consuming more effort and time. Therefore, the physical resources would make it more difficult to use touch resources than mental resources, which would have a greater impact on voice modality usage than mental resources.

The results of Study 1 revealed that there were differences in the physical and mental effort required to use each modality by the menu structure. The capacity of physical and mental resources that can be used by each modality is limited by multitasking contexts, and the required resources would increase according to the menu structure and multitasking contexts. Therefore, the influence of the context on the modality selection would be different by the menu structure.

Hypothesis 6: The multitasking contexts would affect physical effort, mental effort, and satisfaction.

H6a: In multitasking contexts, physical and mental efforts would increase, and satisfaction would decrease compared to Baseline.

H6b: There would be an interaction effect between context and modality on physical effort, mental effort, and satisfaction.

H6C: There would be an interaction effect between context and menu structure on physical effort, mental effort, and satisfaction.

Hypothesis 6 was developed based on Hypothesis 5.

10. METHODOLOGY

10.1. Context

In this study, the ratio of voice and touch that users select when using a multimodal system in four contexts (Baseline, Reading, Watching, and Driving) was compared. These contexts were selected depending on the level of physical and mental resource demands. The compositions of four contexts are shown in Figure 10.1 and the detailed descriptions are as follows:

Baseline (low physical & low mental resource): In the baseline context, there were only modality selection tasks without any additional contexts given to users.

Watching (low physical & high mental resource): The watching context consisted of an animated movie ('Sing[1]') which played on a 12.4-inch tablet (Galaxy tab S7 Plus, Samsung) in front of participants. Their hands were free, but the movie continued to play while they performed the modality selection task.

---

[1] Meledandri, C, Healy, J (Producer) & Jennings, G (Director). (2016). *Sing*, 2016, United States: Illumination Entertainment.

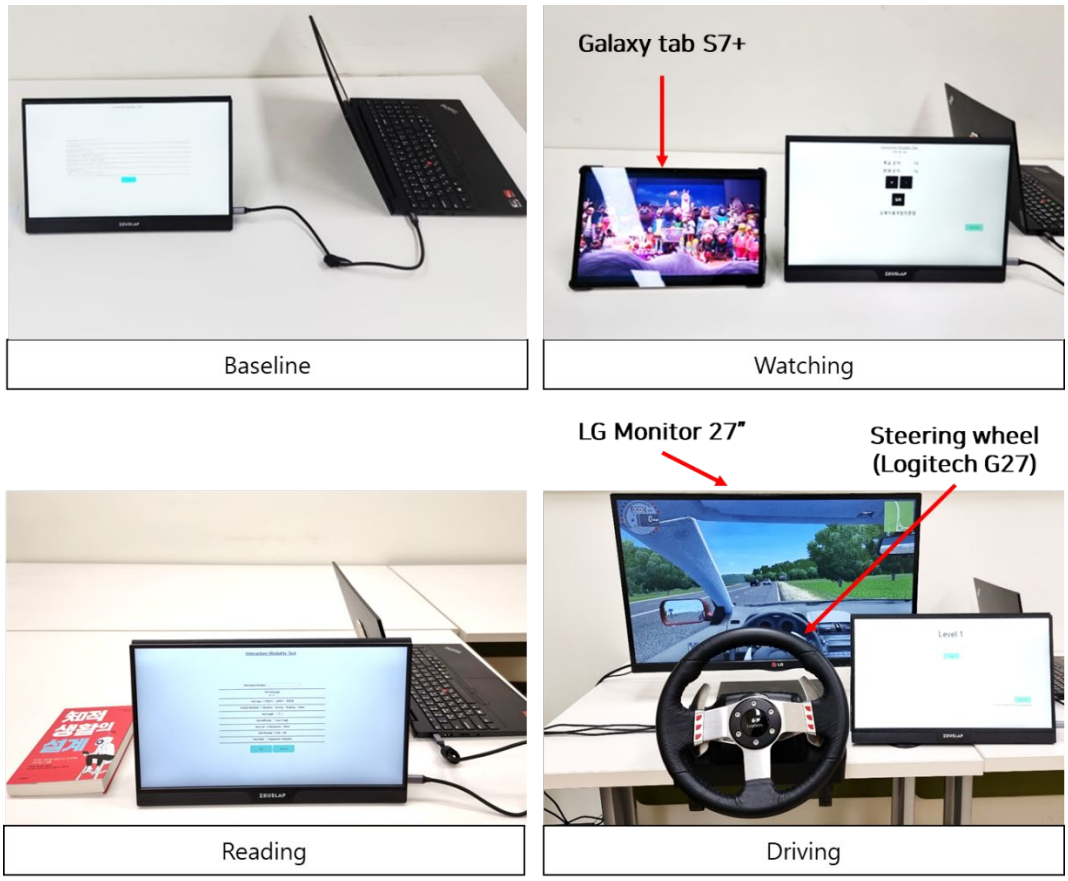


Figure 10.1. Experimental setting for four contexts

Reading (high physical & low mental resource): The reading context consisted of a book that had been selected by participants among five books, including novels and essays. Participants were instructed to hold the book in front of themselves with both hands and read freely at their own pace. They were also instructed to conduct the modality selection task with their right hand only. Tasks were performed on the mobile monitor located on their right side, and after one trial was over, they read the book with both hands again.

Driving (high physical & high mental resource): The driving context consisted of a driving simulator that was equipped with a Logitech® G27 Racing wheel and pedal, and a 27-inch LG monitor. For the driving context, we used the City Car Driving software (Forward Development). A quiet rural highway drive with a total of 6 lanes, 3 for each direction, with a low traffic density (20%) was presented to participants. They were asked to drive an automatic vehicle at an average speed of 60 km/h. Both hands were allowed to hold the handle while driving, and the modality selection task was performed using only their right hand in the same manner as in other contexts.

10.2. Variables

10.2.1. Independent Variables

There were four independent variables: (1) menu structure, (2) number of touches, (3) syllable per touch, and (4) contexts. However, according to study 1, the number of touches was not a significant variable to predict the user's modality selection. Therefore, it was set at the 3-touch which was the middle value in study 1. Contexts was added in as the last variable which was described in section 10.1.

Table 10.1. Independent Variables of study 2

Independent Variable	Level			
Menu Structure	Non-hierarchy		Hierarchy	
Number of Touches	3			
Syllable per Touch (S/T)	1, 2, 3, 4, 5		1, 2, 3, 4, 5, 6, 7, 8	
Context	Baseline	Reading	Watching	Driving

10.2.2. Dependent Variables

The dependent variable was composed of four variables, the same as Study 1.

10.3. Participants

Thirty-one participants (17 males and 14 females) were recruited from the local university. All participants reporting an experience of VUI were Koreans and were between the ages of 22 and 39 (mean age = 27.3 years, SD = 4.19 years). One male participant's baseline data were missing, and one female participant couldn't conduct the reading condition. The experimental procedures, which were also reviewed and approved by the university's Institutional Review Board, were explained to each participant prior to

beginning the study. All participants provided informed consent and received about 20$ for participation.

10.4. Apparatus and Settings

Except for the monitor, steering wheel, tablet, and books for contexts mentioned in section 10.1, the experiment was conducted with same experimental equipment as study 1.

10.5. Procedure

Figure 10.2 shows the overall procedure of study 2. Before the experiment began, the participants read the explanation of the experiment and wrote the consent form consensus. After that, specific instruction on the task was explained. All participants first conducted baseline context as a practice. After that, the remaining three contexts (watching, reading, and driving) were conducted with the void learning effects randomized order, and the order of the menu structure was performed in the order of hierarchy after non-hierarchy. The order of the presented syllable length was balanced and maintained for one level, and there were 5 or 8 levels depending on the menu structure. Each level consisted of 6 trials including the first speech training trial. When the participants completed 6 trials in a certain level, SEQ-based questionnaire items were displayed to evaluate the physical and mental

effort for operating each modality. The effect of the voice modal was measured at all levels, but the effect of the touch modal was measured at random three times in non-hierarchy and four times in hierarchy. This is because the amount of effort required theoretically is the same regardless of the level because it had a constant number of touches regardless of the defect level of the touch modal.

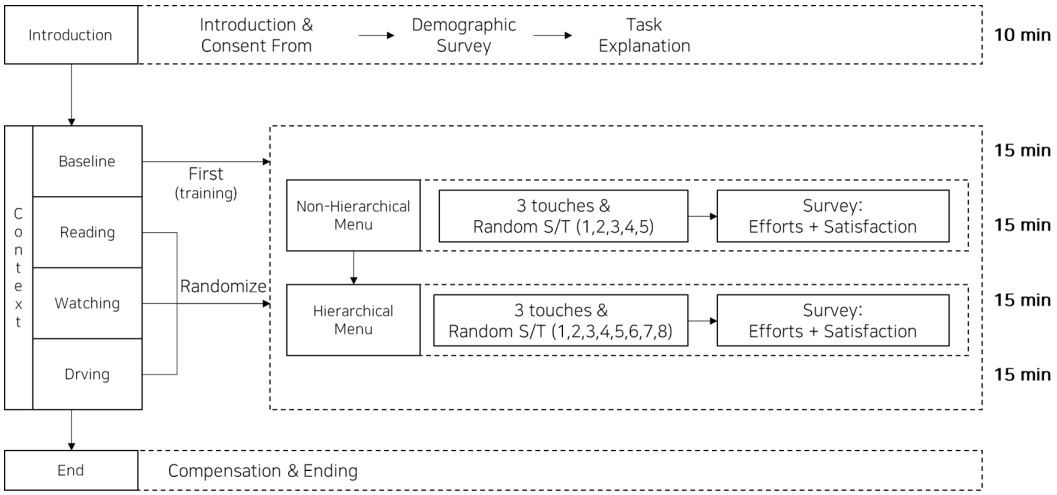


Figure 10.2. Overall procedure of study 2

10.6. Reproducibility of Study

Reproducibility means that the experiment could obtain identical results to a prior study using the same research methods. It is related to whether the experiment was conducted fairly, correctly, and reliably. Therefore, before the main analysis, a simple analysis was conducted using data from both studies to achieve reproducibility.

Data under the same experimental conditions were extracted from studies 1 and 2. Data from the 3-touch condition in study 1 and data from the baseline condition in study 2 were used for analysis. After creating the 'Study' variable, binomial logistic regression analysis was performed using it together with the existing variables, syllable per touch and menu structure. Table X showed the result of logistic regression.

The results revealed that 'study' was not a significant predictor of the regression model (p = .109). This meant that study 2 did not differ from study 1. Even though S/T was presented randomly in study 2, it showed the same results as study 1. On the other hand, S/T and menu structure still significantly predicted the user's modality selection. It was found that, holding all other predictor variables constant, the odds of voice usage decreased by 55% (p = .000, OR: 0.450, 95% CI: 0.427 – 0.473) for each additional increase of 1 S/T. Using the multimodal system in the hierarchy structure was 5.049 times more likely to use the voice modality than in the non-hierarchy structure (p = .000, OR: 5.049, 95% CI: 4.197 – 6.074).

Table 10.2. Result of logistic regression analysis with study 1 and 2 as IV

IV	B	SE	Wald	*p*	OR	95% CI	
						Lower	Upper
Syllable per Touch (S/T)	-0.799	0.026	923.874	.000***	0.45	0.427	0.473
Menu Structure[a] Hierarchy	1.619	0.058	294.924	.000***	5.049	4.197	6.074
Study	-	-	-	.109	-	-	-

reference group: [a]Non-hierarchy

The predicted voice usage of the regression model using study 2 data only was shown in Figure 10.3. The modality switching points of both menu structures were almost the same as in study 1. The modality switching also seems to occur at 2~3 S/T in non-hierarchy and at 4~5 S/T in hierarchy. The percentage correction between the model's prediction and the collected data was 78.2%. The Area Under the Curve (AUC) of the model derived from the ROC curve was 0.849, which means excellent discriminating ability. Therefore, in the consideration of the regression results and modality switching points, the experimental method of this study seems to have obtain reproducibility.

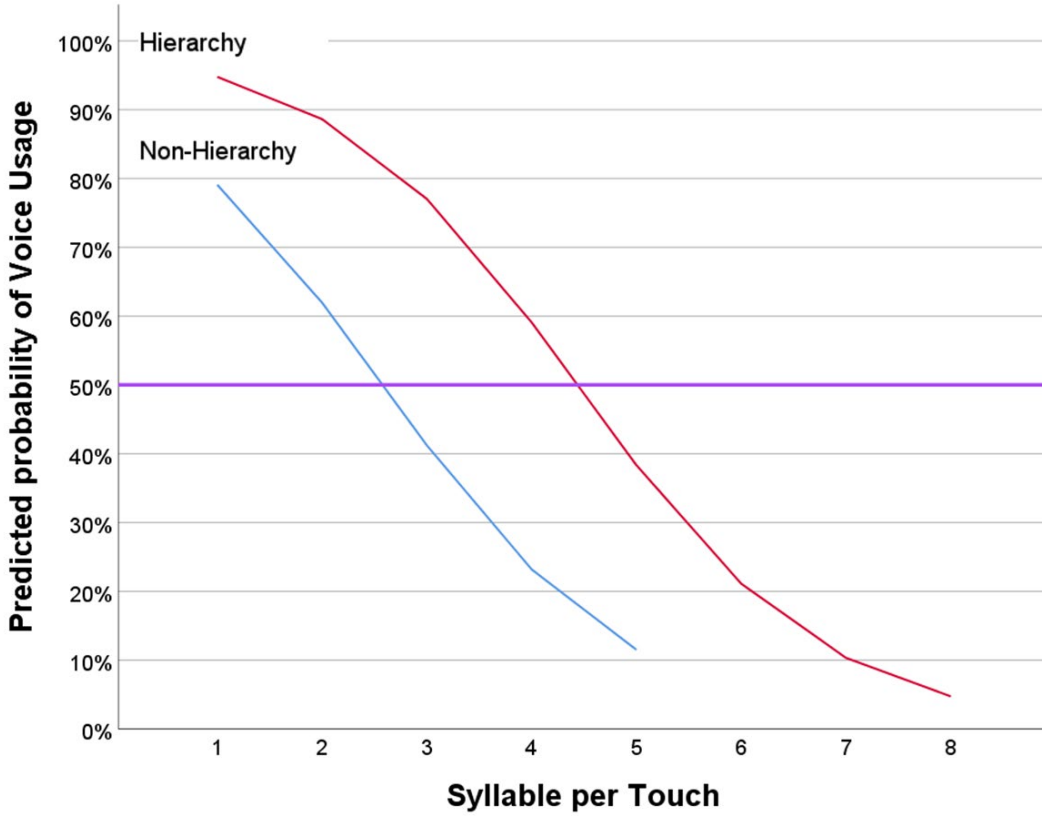


Figure 10.3. Probability of voice usage predicted by logistic regression model with study 2

11. RESULTS

The experimental results were analyzed by menu structure, and 9 data that caused errors during task recording were removed from the analysis. Including two people's data with only three contexts recorded 3043 lines for non-hierarchy conditions and 4878 lines for hierarchy conditions were used for analysis.

11.1. Modality Selection

11.1.1. Non-hierarchy

Since the number of touches was fixed at 3, an analysis was conducted to explore the change in modality usage according to S/T and contexts. Table 11.1 shows the usage of touch and voice modalities according to S/T and contexts. In all contexts, the usage of voice modality increased as S/T decreased. The modality switching point varied depending on the contexts, but it could be predicted that the points were within the conditions of this experiment.

Table 11.1. Descriptive statistics of voice and touch usage in non-hierarchy

Context	Modality	Syllables per Touch (S/T)					Total
		1	2	3	4	5	
Baseline	Touch	30 (20.0%)	54 (36.0%)	98 (65.8%)	121 (80.7%)	121 (80.7%)	424 (56.6%)
	Voice	120 (80.0%)	96 (64.0%)	51 (34.2%)	29 (19.3%)	29 (19.3%)	325 (43.4%)
Watching	Touch	17 (11.1%)	54 (35.1%)	81 (52.3%)	117 (75.5%)	121 (78.1%)	390 (50.5%)
	Voice	136 (88.9%)	100 (64.9%)	74 (47.7%)	38 (24.5%)	34 (21.9%)	382 (49.5%)
Reading	Touch	26 (17.4%)	26 (17.4%)	47 (31.3%)	97 (64.7%)	115 (76.7%)	311 (41.6%)
	Voice	123 (82.6%)	123 (82.6%)	103 (68.7%)	53 (35.3%)	35 (23.3%)	437 (58.4%)
Driving	Touch	5 (3.2%)	10 (6.5%)	38 (24.5%)	64 (41.3%)	86 (55.8%)	203 (26.2%)
	Voice	150 (96.8%)	145 (93.5%)	117 (75.5%)	91 (58.7%)	68 (44.2%)	571 (73.8%)

The binomial logistic regression with forward selection (LR) was conducted to analyze the relationship between S/T, user context, and modality usage. The results are summarized in Table 11.2. As the result of the analysis, S/T and all contexts were found to be significant variables predicting the usage of voice modality at the significance level of 0.05. It was found that, if the context is constant, each additional increase of 1 S/T is associated with a 56/% decrease in the odds of using voice modality ($p < 0.001$, OR: 0.440; 95% CI: 0.411-0.471). When driving, reading, or watching a video, people use voice modality 5.381 times ($p < 0.001$, 95% CI: 4.189-6.912), 2.231 times ($p < 0.001$, 95% CI: 1.761 - 2.828), 1.39 times ($p = 0.005$, 95% CI: 1.103 – 1.754) more than baseline, respectively.

Table 11.2. Summary of logistic regression model of the modality usage in non-hierarchy.

IV		B	SE	Wald	*p*	OR	95% CI	
							Lower	Upper
Syllable per Touch		-0.821	0.034	570.725	0.000***	0.440	0.411	0.471
Context[a]	**Driving**	1.683	0.128	173.581	0.000***	5.381	4.189	6.912
	Reading	0.803	0.121	44.137	0.000***	2.231	1.761	2.828
	Watching	0.329	0.118	7.735	0.005**	1.390	1.102	1.754

**p<.01, ***p<.001
[a]Reference group: Context*Baseline

The predicted usage of voice modality by the regression model was shown in Figure 11.1. The modality switching point, which is the point at which voice and touch usage is reversed, was marked with a purple 50% auxiliary line. The modality switching occurred between 2 and 3 S/T in the baseline context. In the watching context, the modality switching point was almost at 3 S/T, which was higher than in the baseline context. The reading context had the modality switching point at 3~4 S/T and the driving context had it at 4~5 S/T. The percentage correction of the model's used modality classified from the collected data is 74.3%. The Area Under the Curve (AUC) of the model derived from the ROC curve was 0.801, which mean excellent discriminating ability.

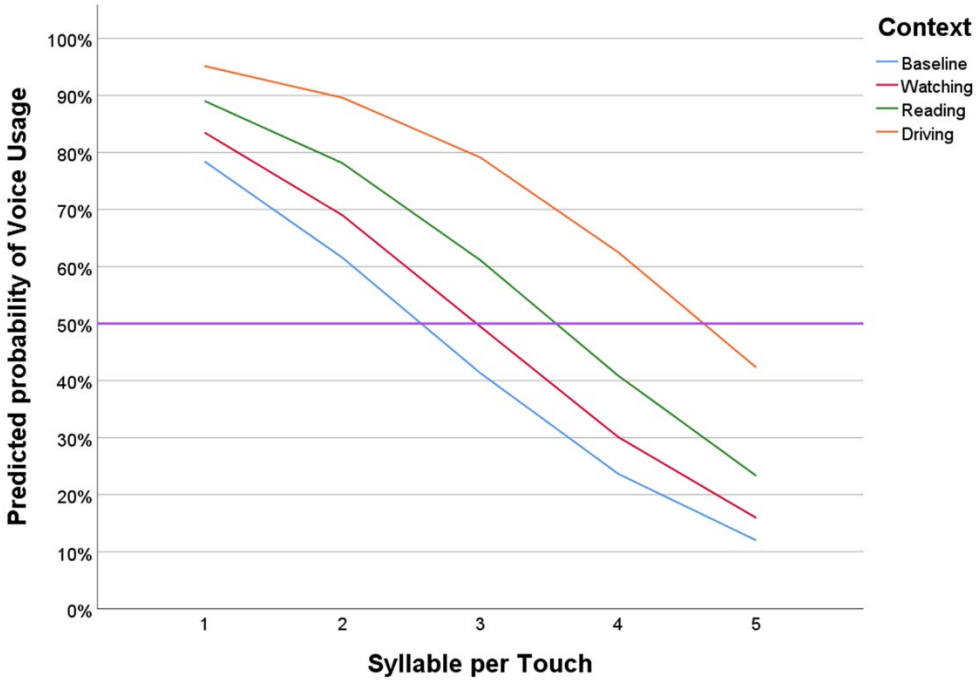


Figure 11.1. Predicted probability of voice usage by contexts in non-hierarchy

11.1.2. Hierarchy

As in non-hierarchy, the effects of S/T and context on modality usage ratio were analyzed. Table 11.3 showed the descriptive statistics of touch and voice modality usage according to S/T and context. In hierarchy modality selection task, the voice modality usage increased as S/T decreased in all contexts. The modality switching points of all contexts could be found within the S/T conditions of this experiment.

The results of binary logistic regression analysis were summarized in Table 11.4. The voice modality usage, holding contexts constant, decreased by 55.8% for every 1 S/T increase ($p < 0.001$, OR: 0.442; 95% CI: 0.423 - 0.462). While driving and reading, participants used the voice modality 4.431 times ($p < 0.001$, 95% CI: 3.547 - 5.535), 1.719 times ($p < 0.001$, 95% CI: 1.388 - 2.128) more than in the baseline, respectively. However, unlike in non-hierarchy, there was no difference in the use of voice modality during watching ($p = 0.608$).

Table 11.3. Descriptive statistics of voice and touch usage in hierarchy

Context	Modality	Syllables per Touch (S/T)								Total
		1	2	3	4	5	6	7	8	
Baseline	Touch	6 (4.0%)	10 (6.7%)	35 (23.3%)	69 (46.0%)	97 (64.7%)	114 (76.0%)	140 (93.3%)	138 (92.0%)	424 (56.6%)
	Voice	144 (96.0%)	140 (93.3%)	115 (76.7%)	81 (54.0%)	53 (35.3%)	36 (24.0%)	10 (6.7%)	12 (8.0%)	325 (43.4%)
Watching	Touch	14 (9.0%)	19 (12.3%)	30 (19.4%)	78 (50.3%)	96 (61.9%)	130 (83.9%)	133 (85.8%)	139 (89.7%)	390 (50.5%)
	Voice	141 (91.0%)	136 (87.7%)	125 (80.6%)	77 (49.7%)	59 (38.1%)	25 (16.1%)	22 (14.2%)	16 (10.3%)	382 (49.5%)
Reading	Touch	0 (0%)	2 (1.3%)	27 (18.0%)	41 (27.3%)	89 (59.3%)	99 (66.0%)	126 (84.0%)	132 (88.6%)	311 (41.6%)
	Voice	150 (100.0%)	148 (98.7%)	123 (82.0%)	109 (72.7%)	61 (40.7%)	51 (34.0%)	24 (16.0%)	17 (11.4%)	437 (58.4%)
Driving	Touch	0 (0%)	1 (0.6%)	2 (1.3%)	27 (17.4%)	58 (37.4%)	77 (50.0%)	100 (64.5%)	108 (69.7%)	203 (26.2%)
	Voice	155 (100.0%)	154 (99.4%)	153 (98.7%)	128 (82.6%)	97 (62.6%)	77 (50.0%)	55 (35.5%)	47 (30.3%)	571 (73.8%)

Table 11.4. Summary of logistic regression model of the modality usage in hierarchy.

IV		B	SE	Wald	*p*	OR	95% CI	
							Lower	Upper
Syllable per Touch		-0.817	0.022	1324.570	0.000***	0.442	0.423	0.462
Context[a]	**Driving**	1.489	0.114	171.864	0.000***	4.431	3.547	5.535
	Reading	0.541	0.109	24.625	0.000***	1.719	1.388	2.128
	Watching	-0.055	0.107	0.263	0.608	0.946	0.767	1.168

***p<.001
[a]Reference group: Context*Baseline

The graph of the voice modality usage predicted by the logistic regression model was displayed in Figure 11.2. In all contexts, the modality switching points of hierarchy conditions moved to the right of those of non-hierarchy conditions. It implied that participants used the voice modality rather than the touch modality despite having to say longer items than non-hierarchy. Baseline and watching had modality switching points between 4 and 5 S/T. the modality switching in reading occurred between 5 and 6 S/T, and in driving occurred between 6 and 7 S/T. The model classified modality accuracy is 79.3%, and the AUC of the ROC curve is 0.879, which means excellent discriminating ability.

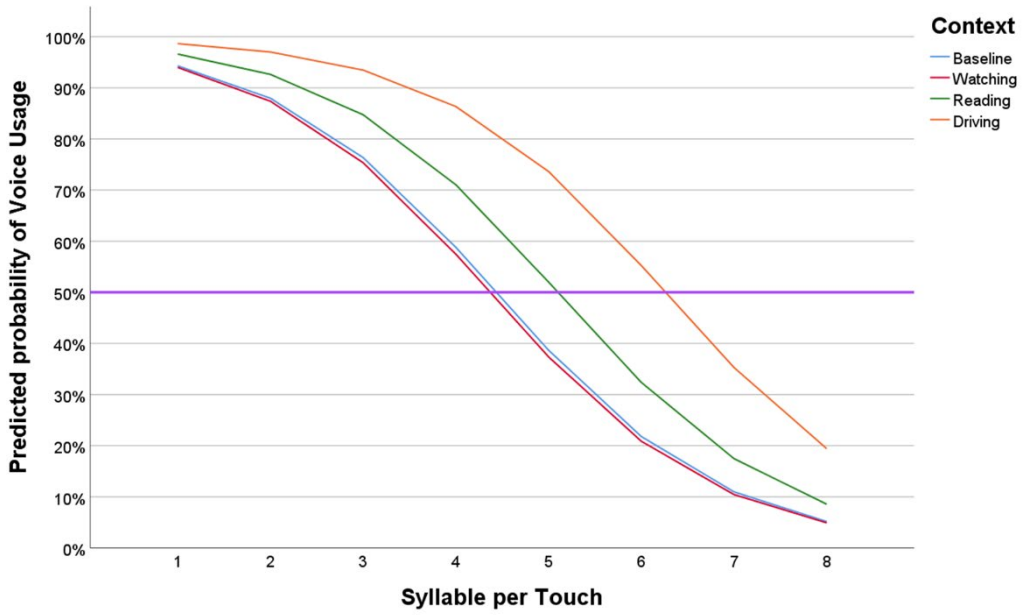


Figure 11.2. Predicted probability of voice usage by contexts in hierarchy

11.2. Physical Effort

Three-way ANOVA was performed to figure out changes in physical effort depending on menu structure, context, and used modality. The results of ANOVA were summarized in Table 11.5. The ANOVA results with physical effort showed that there were significant main effects of menu structure and context [menu structure: $F(1, 1814) = 10.292$, $p = 0.001$; context: $F(3, 1814) = 40.845$, $p < 0.001$; modality: $F(1, 1814) = 1.232$, $p = 0.267$], and only the interaction effect between context and modality was significant among the interaction effects [menu structure × context: $F(3, 1814) = 0.358$, $p = 0.783$; menu structure × modality: $F(1, 1814) = 0.046$, $p = 0.830$; context × modality: $F(3, 1814) = 4.339$, $p = 0.005$; menu structure × context × modality: $F(3, 1814) = 0.147$, $p = 0.932$]. The physical efforts according to the context and modality of each structure were shown in Figure 11.3.

Table 11.5. Result of three-way ANOVA for physical effort by menu structure, context, and modality

DV	Variables	*df*	*F*	*p*	η^2
Physical Effort	Menu Structure	1	10.292	0.001**	0.006
	Context	3	40.845	0.000***	0.063
	Modality	1	1.232	0.267	0.001
	Menu Structure × Context	3	0.358	0.783	0.001
	Menu Structure × Modality	1	0.046	0.830	0.000
	Context × Modality	3	4.339	0.005**	0.007
	Menu Structure × Context × Modality	3	0.147	0.932	0.000

**$p < 0.01$, ***$p < 0.001$

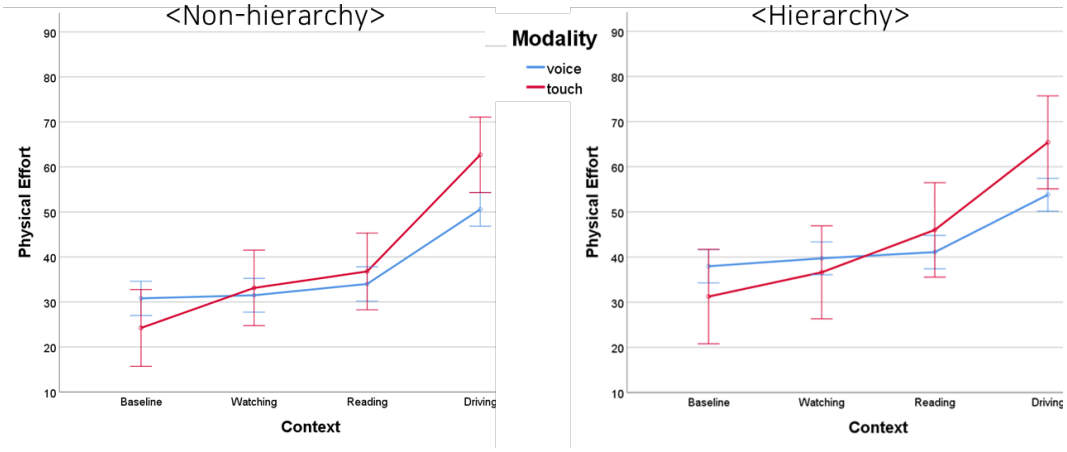


Figure 11.3. Differences in physical effort by modality according to context in each menu structure

Because the interaction effect between context and modality was significant, a simple main effect analysis and a post-hoc analysis were used to analyze the main effect of both variables. Menu structures (non-hierarchy and hierarchy) were analyzed separately.

First, in the non-hierarchy condition, there were significant effects of contexts on physical efforts in both modalities [voice: $F(3, 724) = 23.777$, $p < 0.001$; touch: $F(3, 724) = 14.806$, $p < 0.001$]. However, the Bonferroni pairwise comparison showed that only the driving context was significantly different from other contexts ($ps < 0.000$). In terms of modality, there was a significant effect of modality in the driving context only [baseline: $F(1, 724) = 1.910$, $p = 0.167$; watching: $F(1, 724) = 0.122$, $p = 0.727$; Reading: $F(1, 724) = 0.341$, $p = 0.560$; driving: $F(1, 724) = 6.688$, $p = 0.010$].

In the hierarchy condition, there were also significant effects of contexts on physical efforts in both modalities [voice: $F(3, 1090) = 15.046$, $p < 0.001$; touch: $F(3, 1090) = 8.125$, $p < 0.001$]. And there were significant differences between the driving context and other contexts ($ps < 0.000$). The main effect of modality in context was similar to the results of non-hierarchy condition [baseline: $F(1, 1090) = 1.407$, $p = 0.236$; watching: $F(1, 1090) = 0.312$, $p = 0.577$; Reading: $F(1, 1090) = 0.756$, $p = 0.385$; driving: $F(1, 724) = 4.358$, $p = 0.037$].

11.3. Mental Effort

Three-way ANOVA with mental effort was also performed. The results of ANOVA were summarized in Table 11.6. Most of the results were similar to those of physical effort. The results showed that there were significant main effects of menu structure and context [menu structure: $F(1, 1814) = 13.683$, $p < 0.001$; context: $F(3, 1814) = 60.866$, $p < 0.001$; modality: $F(1, 1814) = 0.594$, $p = 0.441$], and only the interaction effect between context and modality was significant among the interaction effects [menu structure × context: $F(3, 1814) = 0.036$, $p = 0.991$; menu structure × modality: $F(1, 1814) = 0.020$, $p = 0.888$; context × modality: $F(3, 1814) = 3.531$, $p = 0.014$; menu structure × context × modality: $F(3, 1814) = 0.322$, $p = 0.809$]. The mental efforts according to the context and modality of each structure were shown in Figure 11.4.

Table 11.6. Result of three-way ANOVA for mental effort by menu structure, context, and modality

DV	Variables	*df*	*F*	*p*	η^2
Mental Effort	Menu Structure	1	13.683	0.000***	0.007
	Context	3	60.866	0.000***	0.091
	Modality	1	0.594	0.441	0.000
	Menu Structure × Context	3	0.036	0.991	0.000
	Menu Structure × Modality	1	0.020	0.888	0.000
	Context × Modality	3	3.531	0.014*	0.006
	Menu Structure × Context × Modality	3	0.322	0.809	0.001

*$p < 0.05$, ***$p < 0.001$

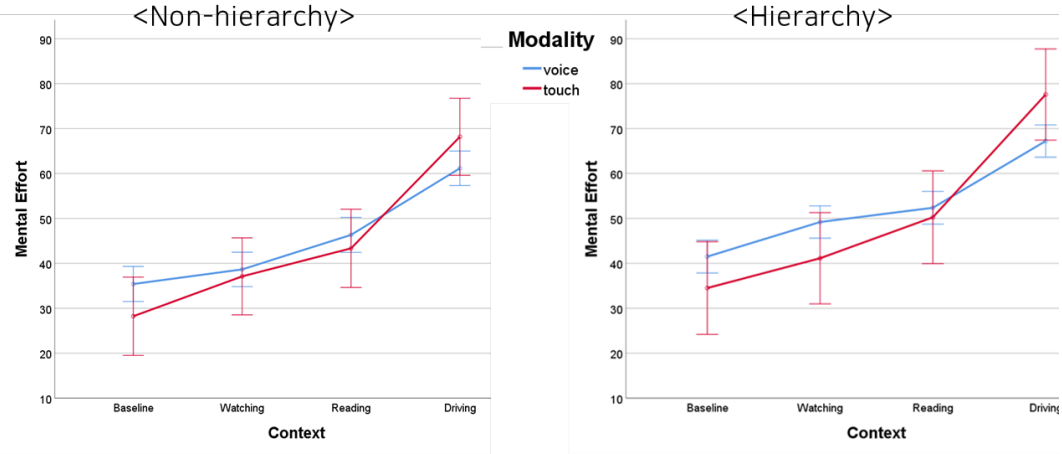


Figure 11.4. Differences in mental effort by modality according to context in each menu structure

Like the result of physical effort, there was a significant interaction effect between context and modality, a simple main effect and a post-hoc analysis were performed to analyze the main effect of both variables.

First, in the non-hierarchy condition, there were significant effects of contexts on physical efforts in both modalities [voice: $F(3, 724) = 34.246$ $p < 0.001$; touch: $F(3, 724) = 15.264$, $p < 0.001$]. Bonferroni pairwise comparison showed that the mental efforts of the baseline context and the watching context did not differ from each other in voice modality ($p = 1.000$). The mental effort of reading context was higher than that of baseline and watching ($p_{r\text{-}b} = 0.001$, $p_{r\text{-}w} = 0.035$). The driving context scored the highest mental effort ($ps < 0.000$). However, in touch modality, the driving context was only significantly different from other contexts ($ps < 0.000$). In terms of modality, the modality did not have a significant effect on mental effort in every context.

In the hierarchy condition, there were also significant effects of contexts on physical efforts in both modalities [voice: $F(3, 1090) = 34.292$, $p < 0.001$; touch: $F(3, 1090) = 13.286$, $p < 0.001$]. Mental effort of voice modality increased: baseline, watching = reading, and driving from lowest to highest. However, in touch modality, the driving context was only significantly different from other contexts ($ps < 0.000$). In terms of modality, the modality did not have a significant effect on mental effort in every context.

11.4. Satisfaction

Three-way ANOVA with satisfaction was also performed. The results of ANOVA were summarized in Table 11.7. Unlike interaction efforts, there was no main effect of menu structure, but the main effects of context and modality were significant. context [menu structure: $F(1, 1814) = 2.171$, $p < 0.141$; context: $F(3, 1814) = 3.489$, $p < 0.015$; modality: $F(1, 1814) = 6.324$, $p = 0.012$]. Only the interaction effect between context and modality was significant among the interaction effects [menu structure × context: $F(3, 1814) = 0.121$, $p = 0.948$; menu structure × modality: $F(1, 1814) = 0.261$, $p = 0.609$; context × modality: $F(3, 1814) = 6.007$, $p < 0.001$; menu structure × context × modality: $F(3, 1814) = 0.214$, $p = 0.887$]. The satisfaction according to the context and modality of each structure were shown in Figure 11.5.

Table 11.7. Result of three-way ANOVA for satisfaction by menu structure, context, and modality

DV	Variables	*df*	*F*	*p*	η^2
Satisfaction	Menu Structure	1	2.171	0.141	0.001
	Context	3	3.489	0.015*	0.006
	Modality	1	6.324	0.012*	0.003
	Menu Structure × Context	3	0.121	0.948	0.000
	Menu Structure × Modality	1	0.261	0.609	0.000
	Context × Modality	3	6.007	0.000***	0.010
	Menu Structure × Context × Modality	3	0.214	0.887	0.000

*$p < 0.05$, ***$p < 0.001$

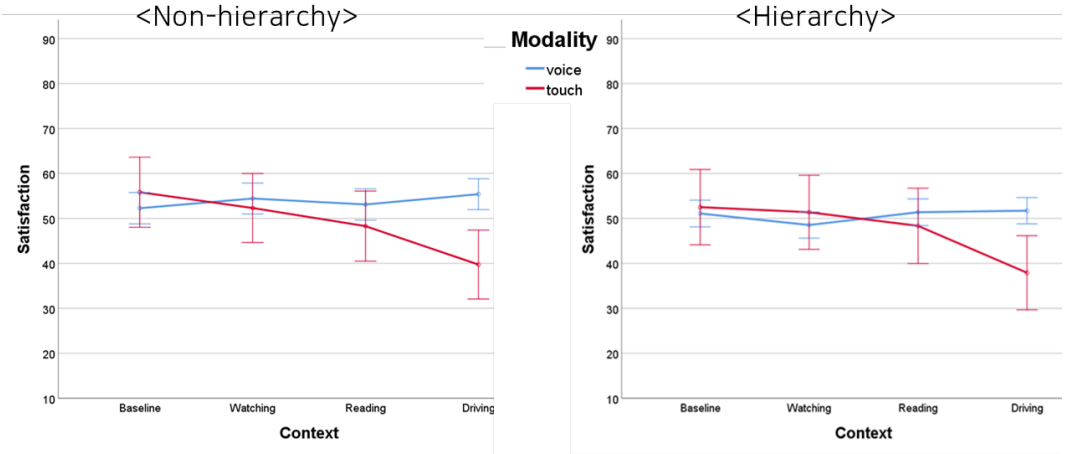


Figure 11.5. Differences in satisfaction by modality according to context in each menu structure

Like the results of interaction efforts, there was a significant interaction effect between context and modality, a simple main effect and post-hoc analysis were performed to analyze the main effect of both variables.

First, in the non-hierarchy condition, the effect of context was significant only in touch modality not in voice modality [voice: $F(3, 724) = 0.616$, $p = 0.605$; touch: $F(3, 724) = 3.110$, $p = 0.026$]. The pairwise comparison showed that there was a significant difference between the baseline and the driving only ($p = 0.024$). In terms of modality, there was a significant effect of modality in the driving context only [baseline: $F(1, 724) = 0.671$, $p = 0.413$; watching: $F(1, 724) = 0.242$, $p = 0.623$; Reading: $F(1, 724) = 1.227$, $p = 0.268$; driving: $F(1, 724) = 13.365$, $p < 0.001$].

In the hierarchy condition, there were no significant effects of context in both modalities. However, there was a significant effect of modality in the driving context ($F(1, 1090) = 9.599$, $p = 0.002$].

12. DISCUSSION

It was found that the results were not significantly different from Study 1, except for the effect of contexts. The reproducibility of the experiment was demonstrated by those results, this dissertation and the experimental method can be said reliable. Syllable per touch and menu structure served as significant predictors of users' modality selection, as in Study 1.

In particular, in Study 2, even though S/T was presented randomly, participants selected the modality according to the change of S/T in various contexts. Participants performed the tasks by using voice modality more than by using touch modality, as syllable per touch decreased. And they used voice modality 5.049 times more in hierarchy than in non-hierarchy.

According to the results of this study, it was confirmed that the user's modality selection is influenced by multitasking contexts. When people use a multimodal system in multitasking situations, they tend to use voice modality more than without context. In non-hierarchy condition, voice modality usage increased in the order of baseline, watching, reading, and driving. On the other hand, in hierarchy condition, voice modality usage also increased, but there was no significant difference between the baseline and the watching.

These results imply that the modality selection between voice and touch is more influenced by the occupancy of physical resources than that of mental resources. As reviewed earlier, this difference is due to the nature of physical and mental resources. Mental resources can be processed in parallel, whereas physical resources can be performed only one action at a time. Therefore, in the reading condition, since the physical effort to the operate touch modality overlapped, it seems that the voice modality usage, which requires less physical effort, was higher. However, just because mental effort can be processed in parallel does not mean that there has been no change in modality selection in the watching condition. These will be explained later through the analysis of interaction

efforts. In addition, the result of the driving context suggests that there was an interaction effect between physical and mental resources.

The reason for the result can be inferred from the change in interaction efforts by contexts. First, in terms of physical effort, there was an interaction effect of context and modality. Although not all contexts had statistically significant differences, physical effort of touch modality increased gradually. However, physical effort of voice modality increased only in the driving context. Therefore, the physical effort of each modality crossed, and this difference might have provided some of the reasons for the modality selection result according to the contexts.

Second, there was also an interaction effect of context and modality in mental effort. Contrary to physical effort, mental effort of voice modality was gradually raised by contexts. In addition, when the mental effort of voice goes from the baseline to the watching, the slope of increase was steeper in hierarchy than in non-hierarchy, and accordingly, the gap with touch modality was not narrowed. In the case of hierarchy, since mental effort required to perform the touch modality was already high, the mental effort increased by watching the video would have been small. Therefore, it is presumed that the reason why there was no difference in modality selection between the baseline and the watching in hierarchy was caused by this difference in the slope of mental effort increase.

Finally, in the result of satisfaction, it was found that satisfaction of voice modality was constant regardless of the context. On the other hand, that of touch steadily decreased.

Participants felt that satisfaction with touch modality was getting poor as the interaction efforts increased, and they began to use voice modality more and more. Interestingly, satisfaction with voice modality did not change even with increased interaction efforts. This seems to be because they evaluated the satisfaction with voice modality by comparing it with touch modality.

The multitasking context study in this chapter highlights the effect of physical and mental resources used in various contexts on modality selection. To understand which modality users select, it is necessary to understand not only the syllable per touch and menu structure revealed in Study 1, but also the context in which users perform tasks. These contexts have interaction effects when physical and mental resource demands are complexly constructed.

CHAPTER V: GENERAL CONCLUSION

13. OVERALL SUMMARY

13.1. Summary of Study 1

Study 1 investigated the effects of modality features (number of touches & syllable per touch) and menu structure on the user's modality selection in a multimodal system. Which modality users would choose was influenced only by syllable per touch, which is the ratio between operating units of modality, rather than individual features of each modality. This is in line with the previous studies that evaluated the performance of individual modalities by using the unit of manipulation (Card et al., 1980; Lee et al., 2019). Therefore, this study can be said to be a more advanced study by expanding the existing studies to the area of comparing the two modalities.

In addition, since the interaction efforts required to manipulate the modalities vary depending on the structures in which functions were composed, the user's modality selection also differed according to menu structure of the multimodal system. This study defined the point at which voice modality starts to be used more than touch modality as 'Modality Switching Point'. There was a modality switching point at less than 3 syllables per touch in non-hierarchy and less than 5 syllables per touch in hierarchy.

Analysis of interaction efforts and satisfaction helped to understand this modality selection. First, physical and mental effort of each modality was affected differently by menu structure. Voice modality maintained the same interaction efforts, but the interaction efforts of touch modality increased in hierarchy. This is the reason for the change of the modality switching point according to the menu structure. Second, people could feel the changes in interaction efforts and satisfaction depending on the modality features and menu structures, and by comparing them, it was possible to decide which modality to use. The fact that the interaction efforts and satisfaction of each modality intersected at the modality switching point supports this opinion.

13.2. Summary of Study 2

As discussed in study 2, contexts of using a multimodal system can influence the interaction efforts of modality and move the modality switching points to more voice-friendly points. These shifts diverged in various degrees to what human resources the contexts occupied. Limiting the user's physical resources rather than their mental resources leads users to select voice modality more. Touch modality was influenced by both physical and mental resources, but voice modality was relatively influenced by only physical resources.

The effect of context was also different depending on the menu structure. Such differences are based on the gaps in the interaction efforts required to manipulate the

modalities in a certain menu structure. The influence of multitasking contexts may change relatively bigger or smaller due to the magnitude of interaction efforts for performing multimodal tasks. The study suggests that the context of the user who uses the multimodal system should be considered and evaluated in order to understand the modality selection between voice and touch modalities.

14. CONCLUSION

14.1. Conclusion and recommendations

Based on the findings and discussion of this dissertation, some brief recommendations are presented for the design of voice modality in multimodal systems.

- When designing a multimodal system, in order to know whether a user will use a certain modality or not, the interaction unit of each modality should be compared.
- To design the voice modality that is frequently used by users, the voice commands should be designed with less than 3 syllables per touch for non-hierarchical functions and less than 5 syllables per touch for hierarchical functions.
- Designers should be reminded that users can recognize the interaction efforts required for manipulating modalities and they can compare those efforts.

- When designing the voice modality, designers should always consider the context of the users using multimodal systems. And the context should be categorized by the physical and mental resources it requires.

- When users encounter physical limitations (e.g., reading a book or cooking) rather than mental limitations (e.g., watching a video or listening to music), they used the voice modality more.

Research and companies developing voice interfaces or voice assistants are still focusing only on the naturalness of commands for voice modality. For instance, when requesting weather information, the command that leads to higher usage would be “Today weather?”, however, to increase naturalness, existing devices train users to say, "What's the weather today?” (Google). However, according to the results of this study, these instructions can make users avoid using voice modality. Even though it is important to pursue the naturalness of voice commands, the length of voice command should be limited so that it does not exceed the modality switching point.

It also revealed that the modality selected by the user can be changed depending on the context even for the same function. Since multimodal systems with voice modality are used in various contexts, it is necessary to evaluate the modality switching point for each context. This research method proposed in this study can guide increasing the use of voice modality and improving the usability of the multimodal system.

14.2. Contribution

This study has several academic and practical implications.

From an academic perspective, this study is meaningful in that it compared the usage of voice and touch modalities by expanding the traditional research method using the manipulation unit of modality. After the appearance of voice modality, researchers have conducted various studies to evaluate the usability of the voice modality, but only under very limited conditions. Nevertheless, this study found and suggested the quantitative method for evaluating the usage of voice modality in multimodal systems. In addition, comparison and verification of the research results were conducted in various contexts. While securing the reliability and validity of the experiment through these results, this study implies that comparisons between voice and various modalities should be made in various contexts. This study also raises a topic in the field of research on the interaction between users and systems, such as human-computer interaction, and encourages further research.

The findings of this study will have great implications for industries related to devices equipped with voice modality or voice assistants, such as smart speakers, vehicle infotainment, and smartphones. These provide insights into designing the voice modality that users can use more easily. It can explain why the developed voice modality is used or not used and can provide guidelines for how to encourage to use of voice modality. Designers will be able to reconsider in advance the usage and usefulness of the voice

modality they plan to develop, which can help develop a user-friendly voice modality and, consequently, improve the usability of whole systems.

14.3. Limitation and Future studies

Although this dissertation provides meaningful findings, there are several limitations that should be considered in future studies.

First, a study to compare voice modality with other types of touch interactions is needed. In this study, the study on modality selection using single-tap touch modality was conducted. However, there are many gestures for touch interaction, and each gesture requires different interaction efforts. If the functions of the multimodal system are operated by different gestures, the modality switching point will also change. Various touch gestures are used in combination with the functions in real systems, and modality selection tasks based on those touch gestures should be designed and tested in future studies.

Second, in this study, the experiment was conducted considering various multitasking contexts. Nevertheless, users use voice and touch modalities in a much wider range of situations. Therefore, research on modality selection in more diverse situations such as listening to music and cooking should be conducted. Through that research, it will be possible to identify the optimal voice modality design for each context or the optimal modality for each context.

Finally, further studies should consider various languages to compare voice and touch modalities. In Korean, one character represents one syllable, but in other languages, the composition can be different and the attitude of people who accept that language could be different. These linguistic and cultural differences should be considered in future studies.

REFERENCES

Arrabito, G. R., Ho, G., Aghaei, B., Burns, C., & Hou, M. (2015). Sustained Attention in Auditory and Visual Monitoring Tasks: Evaluation of the Administration of a Rest Break or Exogenous Vibrotactile Signals. *Human Factors*, *57*(8), 1403–1416. https://doi.org/10.1177/0018720815598433

Beckers, N., Schreiner, S., Bertrand, P., Reimer, B., Mehler, B., Munger, D., & Dobres, J. (2014). Comparing the demands of destination entry using google glass and the samsung Galaxy S4. *Proceedings of the Human Factors and Ergonomics Society*, *2014-Janua*, 2156–2160. https://doi.org/10.1177/1541931214581453

Beirl, D., Rogers, Y., & Yuill, N. (2019). Using voice assistant skills in family life. *Computer-Supported Collaborative Learning Conference, CSCL*, *1*(May), 96–103.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottrjdge, D. (2018). Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–24. https://doi.org/10.1145/3264901

Bilici, V., Krahmer, E., te Riele, S., & Veldhuis, R. (2000). Preferred modalities in dialogue systems. *6th International Conference on Spoken Language Processing, ICSLP 2000*, *Icslp*, 2–5.

Budiu, R. (2013). *Interaction Cost*. https://www.nngroup.com/articles/interaction-cost-definition/

Cabral, M. C., Morimoto, C. H., & Zuffo, M. K. (2005). On the usability of gesture interfaces in virtual reality environments. *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction - CLIHC '05*, 100–108. https://doi.org/10.1145/1111360.1111370

Card, S. K., Mackinlay, J. D., & Robertson, G. G. (1990). The design space of input devices. *Conference on Human Factors in Computing Systems - Proceedings*, *April*, 117–124. https://doi.org/10.1145/97243.97263

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410. https://doi.org/10.1145/358886.358895

Cha, M. C., Hwangbo, H., Lee, S. C., & Ji, Y. G. (2017). F8-3 The Effects of Smartphone Edge Display on EMG Activity of Thumb Muscles in One-handed Interaction. *The Japanese Journal of ErgonomicsThe Japanese Journal of Ergonomics*, *53*(Supplement2), S672–S675. https://doi.org/10.5100/jje.53.s672

Chao, C. M. (2019). Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model. *Frontiers in Psychology*, *10*(JULY), 1–14. https://doi.org/10.3389/fpsyg.2019.01652

Chen, X., & Tremaine, M. (2006). Patterns of multimodal input usage in non-visual information navigation. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *6*(C), 1–10. https://doi.org/10.1109/HICSS.2006.377

Cherubini, M., Anguera, X., Oliver, N., & de Oliveira, R. (2009). Text versus speech. *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '09*, 1. https://doi.org/10.1145/1613858.1613860

Cho, M., Lee, S. S., & Lee, K. P. (2019). Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference*, 1557–1569. https://doi.org/10.1145/3322276.3322332

Cho, N. H. (2002). *Frequent use in modern Korean Research*. Seoul: National Institute of the Korean Language.

Christie, J., Klein, R. M., & Watters, C. (2004). *A comparison of simple hierarchy and grid metaphors for option layouts on small-size screens*. *60*, 564–584. https://doi.org/10.1016/j.ijhcs.2003.10.003

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). Voice User Interface Design Library of Congress Cataloging-in-Publication Data Voice user interface design. In *Voice User Interface Design* (Issue January). https://doi.org/10.1093/combul/46.5.30-b

Cox, A. L., Cairns, P. A., Walton, A., & Lee, S. (2008). Tlk or txt? Using voice input for SMS composition. *Personal and Ubiquitous Computing*, *12*(8), 567–588. https://doi.org/10.1007/s00779-007-0178-8

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., & Abbott, S. (2003). The development and evaluation of a speech-to-sign translation system to assist transactions. *International Journal of Human-Computer Interaction*, *16*(2), 141–161. https://doi.org/10.1207/S15327590IJHC1602

Du, Y., Qin, J., Zhang, S., Cao, S., & Dou, J. (2018). Voice User Interface Interaction Design Research Based on User Mental Model in Autonomous Vehicle. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9170, pp. 117–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-91250-9_10

Fujimura, O. (1975). Syllable as a Unit of Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(1), 82–87. https://doi.org/10.1109/TASSP.1975.1162631

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482. https://doi.org/10.1037/0033-295X.113.3.461

Google. (n.d.). What you can ask Google Assistant, Retrieved Dec 02, 2022, from https://support.google.com/assistant/answer/7172842

Han, S. H., & Kwahk, J. (1994). *Design of a Menu for Small Displays Presenting a Single Item at a Time*. 360–364. https://doi.org/https://doi.org/10.1177/154193129403800502

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Power Technology and Engineering* (Vol. 43, Issue 5, pp. 139–183). https://doi.org/10.1016/S0166-4115(08)62386-9

He, J., Choi, W., McCarley, J. S., Chaparro, B. S., & Wang, C. (2015). Texting while driving using Google Glass™: Promising but not distraction-free. *Accident Analysis and Prevention*, *81*, 218–229. https://doi.org/10.1016/j.aap.2015.03.033

Hong, J., & Findlater, L. (2018). Identifying speech input errors through audio-only interaction. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–12. https://doi.org/10.1145/3173574.3174141

Hong, L., Chi, E. H., Budiu, R., Pirolli, P., & Nelson, L. (2008). SparTag.us: A low cost tagging system for foraging of web content. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, 65–72. https://doi.org/10.1145/1385569.1385582

Hua, Z., & Ng, W. L. (2010). Speech recognition interface design for in-vehicle system. *Proceedings of the 2nd International Conference on Automotive User Interfaces and*

Interactive Vehicular Applications - AutomotiveUI '10, *AutomotiveUI*, 29. https://doi.org/10.1145/1969773.1969780

Hwangbo, H., Yoon, S. H., Jin, B. S., Han, Y. S., & Ji, Y. G. (2013). A Study of Pointing Performance of Elderly Users on Smartphones. *International Journal of Human-Computer Interaction*, *29*(9), 604–618. https://doi.org/10.1080/10447318.2012.729996

Jameson, A., & Kl¨ockner, K. (2005). USERMULTITASKINGWITHMOBILEMULTIMODAL SYSTEMS. *Spoken Multimodal Human- Computer Dialogue in Mobile Environments*, 349–377.

Jiang, L., & Chen, Y. H. (2022). Menu Design on Small Display User Interfaces: Measuring the Influence of Menu Type, Number of Preview Items, and Menu Breadth on Navigation Efficiency. *International Journal of Human-Computer Interaction*, *38*(7), 631–645. https://doi.org/10.1080/10447318.2021.1954781

Jung, J., Lee, S., Hong, J., Youn, E., & Lee, G. (2020). Voice+Tactile: Augmenting In-vehicle Voice User Interface with Tactile Touchpad Interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. https://doi.org/10.1145/3313831.3376863

Kim, H., & Song, H. (2014). Evaluation of the safety and usability of touch gestures in operating in-vehicle information systems with visual occlusion. *Applied Ergonomics*, *45*(3), 789–798. https://doi.org/10.1016/j.apergo.2013.10.013

Kim, J., Jeong, M., & Lee, S. C. (2019). “Why did this voice agent not understand me?”: Error Recovery Strategy for In-Vehicle Voice User Interface. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct Proceedings - AutomotiveUI ’19*, 146–150. https://doi.org/10.1145/3349263.3351513

Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H., & Temem, J. N. (2002). User evaluation of the Mask kiosk. *Speech Communication*, *38*(1–2), 131–139. https://doi.org/10.1016/S0167-6393(01)00048-6

Laureiti, C., Cordella, F., di Luzio, F. S., Saccucci, S., Davalli, A., Sacchetti, R., & Zollo, L. (2017). Comparative performance analysis of M-IμEMG and voice user interfaces for assistive robots. *IEEE International Conference on Rehabilitation Robotics*, 1001–1006. https://doi.org/10.1109/ICORR.2017.8009380

Lee, S. C., Yoon, S. H., & Ji, Y. G. (2019). Modeling task completion time of in-vehicle information systems while driving with keystroke level modeling. *International Journal of Industrial Ergonomics*, *72*(June), 252–260. https://doi.org/10.1016/j.ergon.2019.06.001

Lemmelä, S. (2008). Selecting optimal modalities for multimodal interaction in mobile and pervasive environments. *Proceedings of IMUx (Improved Mobile User Experience) Workshop*.

Lemmelä, S., Vetek, A., Mäkelä, K., & Trendafilov, D. (2008). Designing and evaluating multimodal interaction for mobile contexts. *ICMI'08: Proceedings of the 10th International Conference on Multimodal Interfaces*, 265–272. https://doi.org/https://doi.org/10.1145/1452392.1452447

Li, J. (2022). Recent Advances in End-to-End Automatic Speech Recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1), 106–124. https://doi.org/10.1561/116.00000050

Li, R., Victor, Y., Sha, C., & Lu, Z. (2017). Effects of interface layout on the usability of In-Vehicle Information Systems and driving safety. *Displays*, *49*, 124–132. https://doi.org/10.1016/j.displa.2017.07.008

Liu, X., & Thomas, G. W. (2017). Gesture interfaces: Minor change in effort, major impact on appeal. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 4278–4283. https://doi.org/10.1145/3025453.3025513

Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 580–628. https://doi.org/10.1145/3025453.3025786

Mane, A., Boyce, S., Karis, D., & Yankelovich, N. (1996). *Designing the user interface for speech recognition applications*. *28*(4), 431. https://doi.org/10.1145/257089.257431

Medhi, I., Toyama, K., Joshi, A., Athavankar, U., & Cutrell, E. (2013). A Comparison of List vs. Hierarchical UIs on Mobile Phones for Non-literate Users. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8118 LNCS* (Issue PART 2, pp. 497–504). https://doi.org/10.1007/978-3-642-40480-1_33

Mitchard, H., & Winkles, J. (2002). Experimental comparisons of data entry by automated speech recognition, keyboard, and mouse. *Human Factors*, *44*(2), 198–209. https://doi.org/10.1518/0018720024497907

Naumann, A. B., Wechsung, I., & Möller, S. (2008). Factors influencing modality choice in multimodal applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5078 LNCS*, 37–43. https://doi.org/10.1007/978-3-540-69369-7_5

Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study. *Computers in Human Behavior*, *128*(May 2021), 107093. https://doi.org/10.1016/j.chb.2021.107093

Nguyen, Q. N., Ta, A., & Prybutok, V. (2019). An Integrated Model of Voice-User Interface Continuance Intention: The Gender Effect. *International Journal of Human-Computer Interaction*, *35*(15), 1362–1377. https://doi.org/10.1080/10447318.2018.1525023

Oviatt, S., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? *Proceedings of the 6th International Conference on Multimodal Interfaces - ICMI '04*, 129. https://doi.org/10.1145/1027933.1027957

Paap, K. R., & Cooke, N. J. (1997). Design of Menus. *Handbook of Human-Computer Interaction*, 533–572. https://doi.org/10.1016/B978-044481862-1.50090-X

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Park, S., Ha, J., & Kim, L. (2021). Anti-Heartbeat-Evoked Potentials Performance in Event-Related Potentials-Based Mental Workload Assessment. *Frontiers in Physiology*, *12*(October). https://doi.org/10.3389/fphys.2021.744071

Pearl, Cathy. (2016). *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media, Inc. https://books.google.co.kr/books?hl=ko&lr=&id=MmnEDQAAQBAJ&oi=fnd&pg=PR11&dq=VUI+advantage+multitasking&ots=HNb-0ubBhd&sig=GkFodoix6KBzwMh-UC_yDYZaBhU#v=onepage&q=VUI advantage multitasking&f=false

Peissner, M., & Doebler, V. (2011). Can voice interaction help reducing the level of distraction and prevent accidents ? *Whitepaper*, *May*, 24.

Perakakis, M., & Potamianos, A. (2008a). A study in efficiency and modality usage in multimodal form filling systems. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(6), 1194–1206. https://doi.org/10.1109/TASL.2008.2001389

Perakakis, M., & Potamianos, A. (2008b). Multimodal system evaluation using modality efficiency and synergy metrics. *Proceedings of the 10th International Conference on Multimodal Interfaces - IMCI '08*, 9. https://doi.org/10.1145/1452392.1452397

Perlman, D., Samost, A., Domel, A. G., Mehler, B., Dobres, J., & Reimer, B. (2019). The relative impact of smartwatch and smartphone use while driving on workload, attention, and driving performance. *Applied Ergonomics*, *75*(June 2017), 8–16. https://doi.org/10.1016/j.apergo.2018.09.001

Perugini, S., Anderson, T. J., & Moroney, W. F. (2007). A Study of Out-of-turn Interaction in Menu-based, IVR, Voicemail Systems. *Proceedings of ACM CHI 2007: Conference on Human Factors in Computing Systems*, 961–970.

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–12. https://doi.org/10.1145/3173574.3174214

Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2013). The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance. *MIT*

AgeLab Technical Report No. 2013-17A. Massachusetts Institute of Technology, Cambridge, MA., *May 2014*, 293. https://doi.org/2013-18A

Resnick, P., & Virzi, R. A. (1992). Skip and scan. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 419–426. https://doi.org/10.1145/142750.142881

Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. (2016). *Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices*.

Rupp, M. A., Michaelis, J. R., McConnell, D. S., & Smither, J. A. (2018). The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact. *Applied Ergonomics*, *70*(April 2017), 77–87. https://doi.org/10.1016/j.apergo.2018.02.005

Saon, G., Tuske, Z., Bolanos, D., & Kingsbury, B. (2021). Advancing RNN Transducer Technology for Speech Recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2*, 5654–5658. https://doi.org/10.1109/ICASSP39728.2021.9414716

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. *Conference on Human Factors in Computing Systems - Proceedings*, 1599–1608. https://doi.org/10.1145/1518701.1518946

Schaffer, S., Jöckel, B., Wechsung, I., Schleicher, R., & Möller, S. (2011). Modality selection and perceived mental effort in a mobile application. *Proceedings of the*

Annual Conference of the International Speech Communication Association, INTERSPEECH, *August*, 2253–2256. https://doi.org/10.21437/interspeech.2011-599

Schwaller, M., & Lalanne, D. (2013). Pointing in the Air: Measuring the Effect of Hand Selection Strategies on Performance and Effort. In *LNCS* (Vol. 7946, pp. 732–747). https://doi.org/10.1007/978-3-642-39062-3_53

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?" *Proceedings of the 2018 Designing Interactive Systems Conference*, 857–868. https://doi.org/10.1145/3196709.3196772

Simpson, G. B., & Kang, H. (2004). Syllable processing in alphabetic Korean. *Reading and Writing*, *17*(1–2), 137–151. https://doi.org/10.1023/b:read.0000013808.65933.a1

Standardization, I. O. for. (1998). *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs): Part 11: Guidance on usability*.

Strayer, D. L., Cooper, J. M., McCarty, M. M., Getty, D. J., Wheatley, C. L., Motzkus, C. J., Goethe, R. M., Biondi, F., & Horrey, W. J. (2019). Visual and Cognitive Demands of CarPlay, Android Auto, and Five Native Infotainment Systems. *Human Factors*, *61*(8), 1371–1386. https://doi.org/10.1177/0018720819836575

Suhm, B., Myers, B., & Waibel, A. (1999). Model-based and empirical evaluation of multimodal interactive error correction. *Conference on Human Factors in Computing Systems - Proceedings*, 584–591. https://doi.org/10.1145/302979.303165

Tsimhoni, O., Smith, D., & Green, P. (2004). Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human Factors*, *46*(4), 600–610. https://doi.org/10.1518/hfes.46.4.600.56813

Turner, C. J., Chaparro, B. S., & He, J. (2020). Typing on a Smartwatch While Mobile: A Comparison of Input Methods. *Human Factors*. https://doi.org/10.1177/0018720819891291

Vailshery, L. S. (2021). *Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)*. Statista.

van Pinxteren, M. M. E., Pluymaekers, M., & Lemmink, J. G. A. M. (2020). Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management*, *31*(2), 203–225. https://doi.org/10.1108/JOSM-06-2019-0175

Voicebot, A. I. (2019). *Smart speaker consumer adoption report*.

Voicebot.ai. (January 14, 2021). Total number of Amazon Alexa skills from January 2019 to January 2021, by country [Graph]. In Statista. Retrieved December 5, 2022, from https://www.statista.com/statistics/1189221/amazon-alexa-total-skills/

Wallace, D. F., Anderson, N. S., & Ben, S. (1987). *Time Stress Effects on Two Menu Selection Systems*. *1986*, 727–731. https://doi.org/10.1177/154193128703100708

Westbrook, R. A. (1980). A Rating Scale for Measuring Product/Service Satisfaction. *Journal of Marketing*, *44*(4), 68. https://doi.org/10.2307/1251232

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). Engineering Psychology and Human Performance. In *Engineering Psychology and Human Performance*. Routledge. https://doi.org/10.4324/9781003177616

Widyanti, A., Johnson, A., & de Waard, D. (2013). Adaptation of the Rating Scale Mental Effort (RSME) for use in Indonesia. *International Journal of Industrial Ergonomics*, *43*(1), 70–76. https://doi.org/10.1016/j.ergon.2012.11.003

Wijnen, F. (1988). Spontaneous word fragmentations in children: evidence for the syllable as a unit in speech production. *Journal of Phonetics*, *16*(2), 187–202. https://doi.org/10.1016/s0095-4470(19)30486-3

Zaphiris, P., Shneiderman, B., & Norman, K. L. (2002). Expandable indexes vs. sequential menus for searching hierarchies on the world wide web. *Behaviour and Information Technology*, *21*(3), 201–207. https://doi.org/10.1080/0144929021000009045

Zijlstra, F. R. (1993). Efficiency in work behaviour: A design approach for modern tools. *Delft University Press*, *January 1993*, 1–186.

http://www.csa.com/partners/viewrecord.php?requester=gs&collection=TRD&recid
=N9516953AH

APPENDIX

Appendix 1. Demographic information of participants in Study 1

No	Age	Gender	Frequency of VUI use[1)]	Used voice assistants or services
P101	28	Male	2	시리, 네비게이션
P102	25	Male	4	빅스비
P103	30	Male	1	빅스비, 시리, 안드로이드오토
P104	26	Male	4	시리
P105	26	Female	3	시리
P106	22	Female	3	시리, 클로바, 지니
P107	37	Male	1	빅스비
P108	34	Male	1	클로바노트, AI 스피커(카카오)
P109	22	Female	1	구글 어시스턴트, 기가지니, 시리, 빅스비
P110	22	Male	4	빅스비
P111	28	Male	1	구글
P112	25	Female	3	시리
P113	26	Male	2	빅스비
P114	23	Male	3	시리
P115	26	Male	4	시리
P116	25	Female	4	시리
P117	37	Female	4	시리
P118	25	Male	2	시리
P119	27	Female	4	시리
P120	39	Male	1	빅스비, 맵피, 구글 어시스턴트
P121	28	Male	3	빅스비, 에이닷, 네비게이션(벤츠)
P122	23	Female	3	시리, 빅스비, 카카오
P123	27	Female	2	시리

P124	34	Female	2	시리, 아리
P125	28	Female	2	티맵, 시리
P126	26	Female	3	시리
P127	26	Female	3	빅스비
P128	27	Male		missing
P129	29	Male		missing
P130	26	Male		protocol violation

[1)] 1: Daily, 2: Weekly, 3: Monthly. 4: Rarely

Appendix 2. Demographic information of participants in Study 2

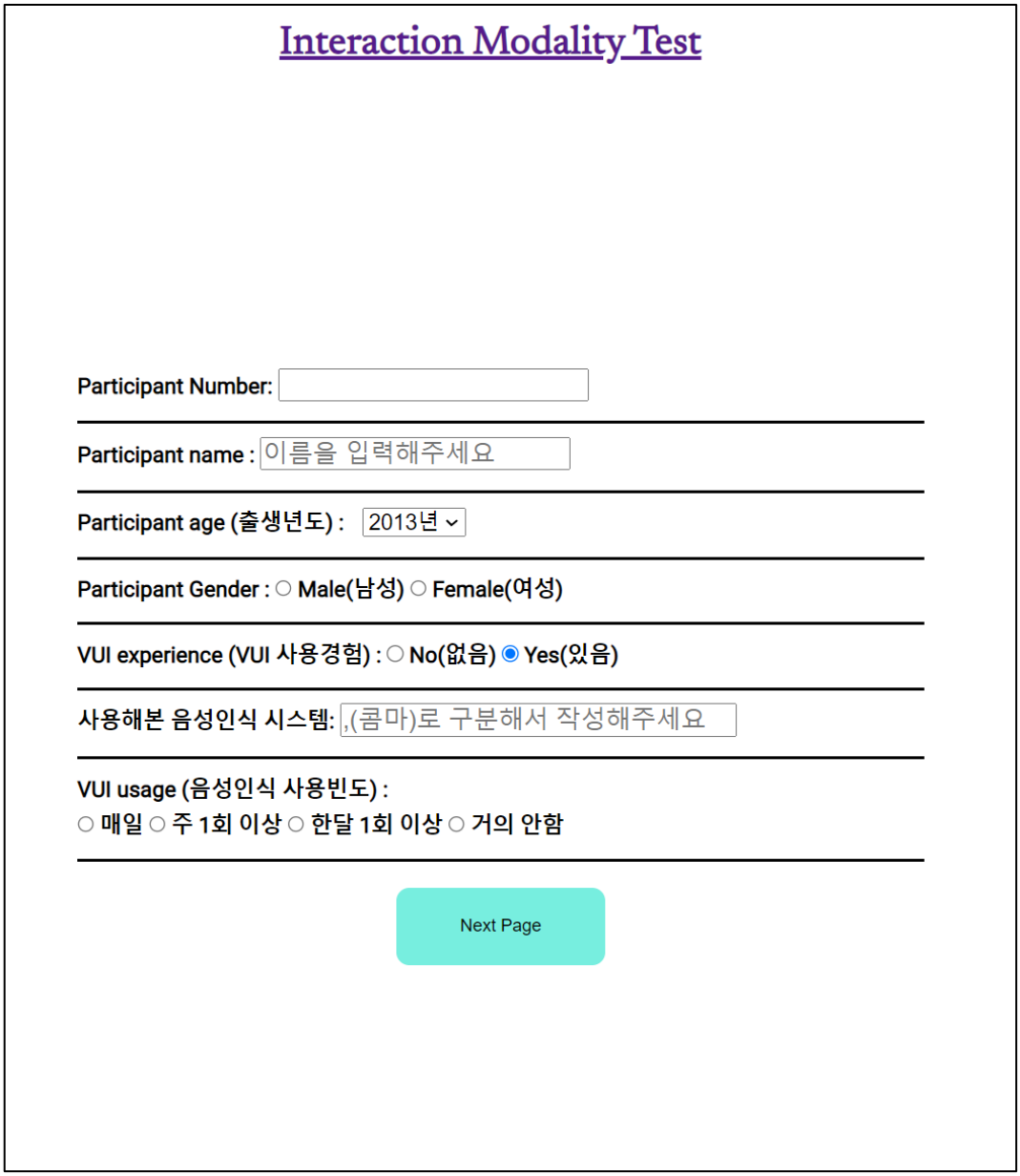
No	Age	Gender	Frequency of VUI use[1)]	Used voice assistants or services
P201	26	Male	4	시리
P202	25	Male	2	시리
P203	39	Male	1	빅스비, 구글 어시스턴트, 맵피
P204	31	Male	3	클로버, 빅스비
P205	31	Male	3	시리, 아리
P206	33	Male	3	티맵
P207	24	Male	3	빅스비, 기가지니
P208	26	Female	2	시리
P209	25	Male	3	빅스비
P210	26	Male	4	시리
P211	23	Male	1	시리, 빅스비
P212	24	Female	2	시리
P213	28	Female	4	시리
P214	28	Male	2	빅스비, 아리
P215	27	Male	4	시리
P216	27	Male	4	시리
P217	26	Female	2	시리, 현대, 빅스비
P218	24	Female	2	시리
P219	25	Female	3	카카오, 시리, 자동차(기아, 현대)
P220	26	Female	1	빅스비, 시리, 카카오
P221	25	Female	2	지니, 빅스비
P222	34	Male	3	시리, 워드, 아리
P223	27	Female	2	시리
P224	25	Female	4	시리

P225	22	Female	1	아리아, 시리, 빅스비, 현대
P226	26	Female	1	시리
P227	23	Female	3	시리, 빅스비, 카카오네비
P228	37	Male	2	시리
P229	31	Male	2	티맵, 빅스비, 클로버
P230	22	Female	4	빅스비
P231	31	Male	1	클로바, 시리

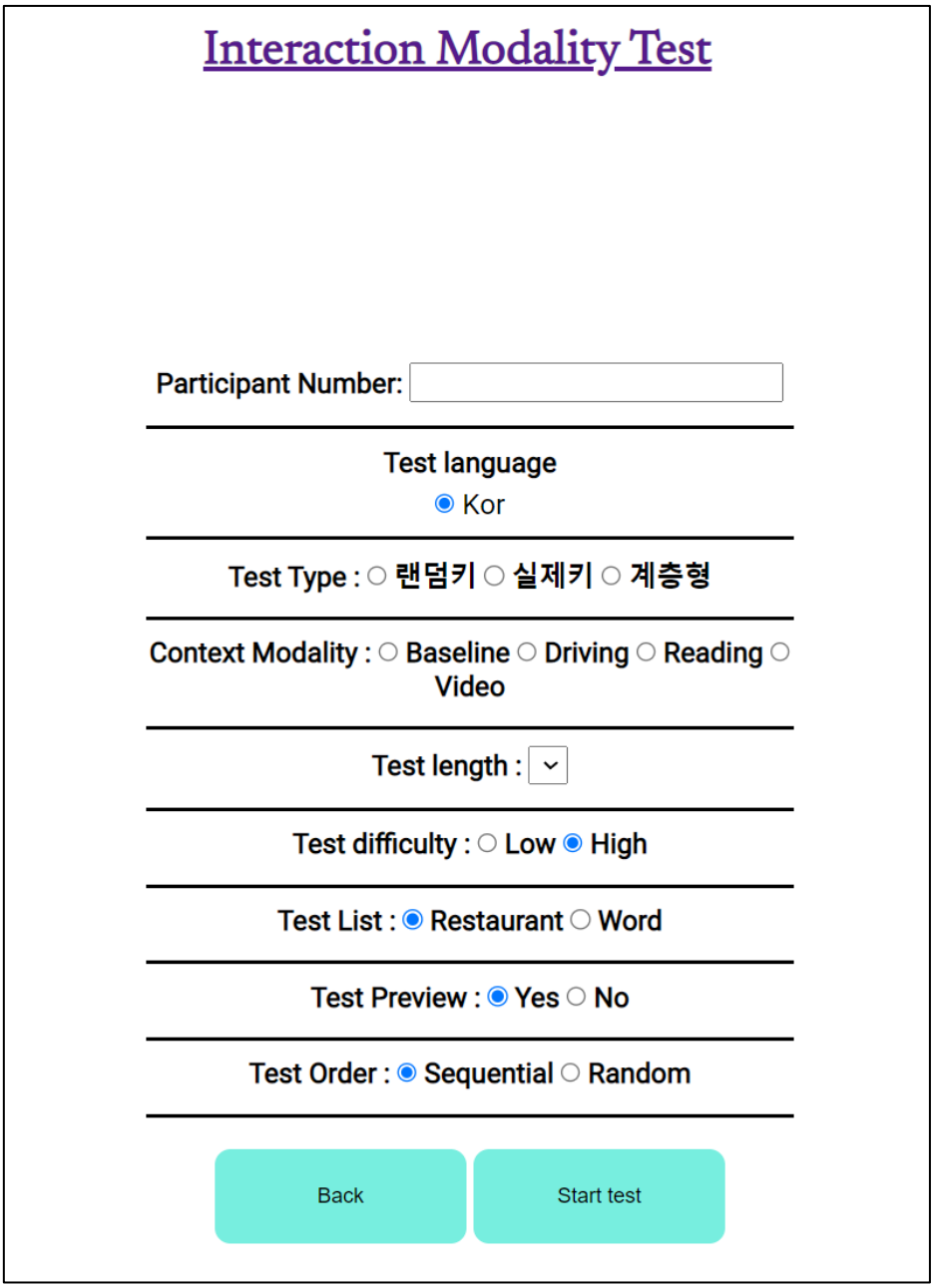
[1)] 1: Daily, 2: Weekly, 3: Monthly. 4: Rarely

Appendix 3. Experimental Program

1. Demographic Survey Page

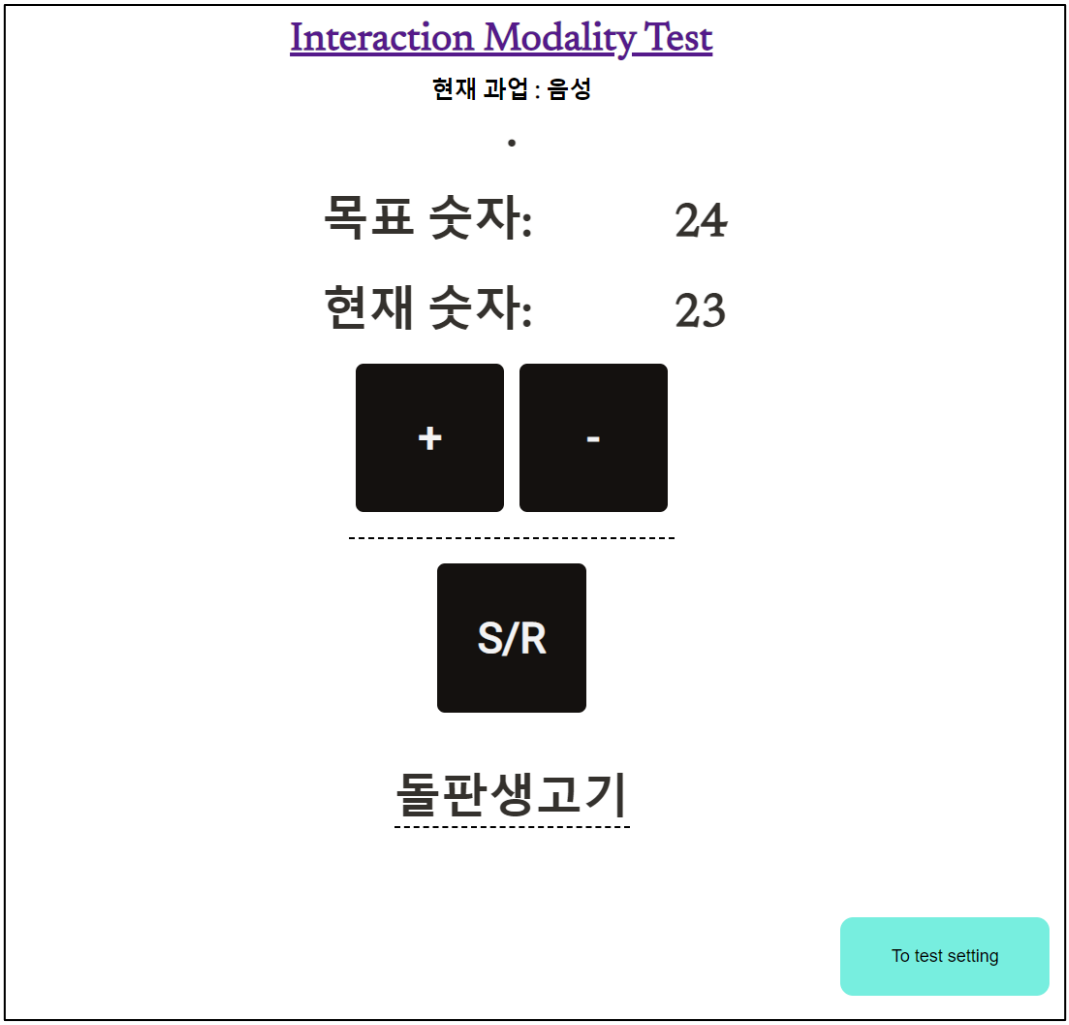


2. Settings Page

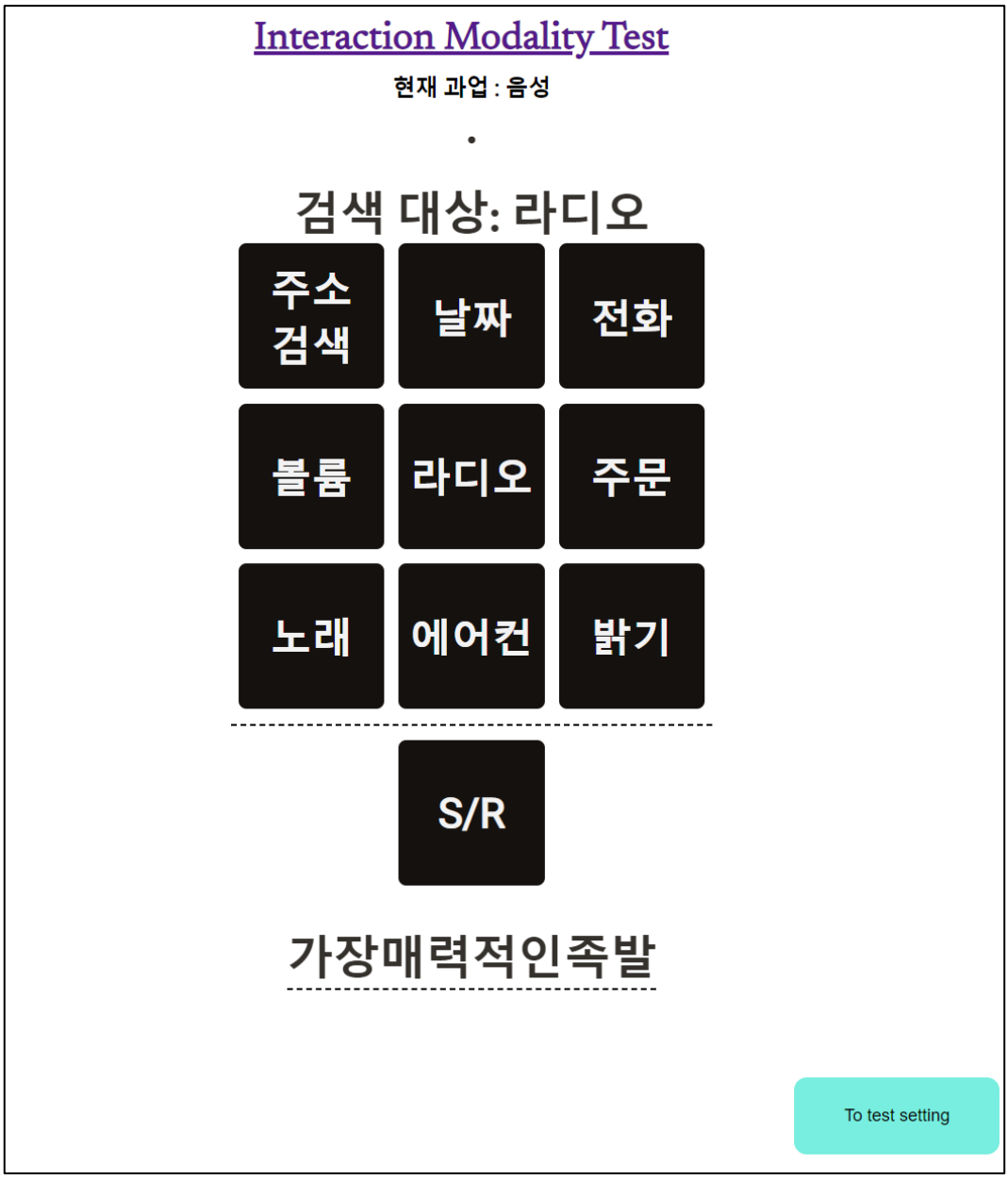


3. Test Ready Page

4. Non-hierarchy Modality Selection Page



5. Hierarchy Modality Selection Page



Appendix 4. Approval from the Yonsei University Institutional Review Board (IRB)

연세대학교
생명윤리위원회

결과통보서

결과통보서

신청번호	202210-HR-3022-02					
연구과제명	상호작용 비용 관점에서의 음성 사용자 인터페이스 조사					
연구책임자	**성명**	차민철	**학과**	정보산업공학과	**직위**	석박사통합과정
접수일자	2022-10-05		**통보일자**	2022-10-11		
심의일자	2022-10-11		**승인번호**	7001988-202210-HR-1709-02		
심의대상	○ 신규 ○ 보완 ○ 종료보고		◉ 수정 후 승인 ○ 변경심의 ○ 결과보고		○ 수정 후 신속심의 ○ 지속심의 ○ 기타보고	
심의종류	○ 정규심의	◉ 신속심의	○ 심의면제			
심의결과	◉ 승인	○ 수정후승인	○ 수정후신속심의	○ 보완	○ 반려	○ 서류보완
연구기간	IRB 승인일 이후 ~ 2023-02-28 (5개월)					
승인유효기간	2022-10-11 ~ 2023-02-28					
연구위험성	○ Level1	◉ Level2	○ Level3	○ Level4		
심의자료	심의의뢰서 , 연구계획서 , 피험자 설명문 , 기타					
심의결과에 대한 사유 및 의견	다음의 수정사항 반영을 확인함. 1. (연구계획서) 문구 수정 2. (연구참여자 설명문) 문구 수정 3. (연구참여자 모집문건) 문구 수정 본 연구는 음성 사용자 인터페이스(VUI)의 상호작용 비용 영향 요인 및 음성 사용자 인터페이스와 터치 인터렉션의 상호작용 비용 및 주관적 작업부하가 모달리티 선택에 미치는 영향을 알아보고자 함. 성인(100명)을 대상으로 컴퓨터 과제수행 및 설문연구(1회, 1시간)를 수행하고자 함. 연구 수행 이전에 충분히 연구에 대한 설명 및 주의사항을 읽고 자발적 동의를 구한 후 연구를 수행하도록 조치하며 및 중도철회, 예측 부작용과 조치 등 연구수행에 있어 과학적, 윤리적으로 잘 보완하고 있으므로 승인하고자 함. * 연구위험성: Level 2(최소위험에서 약간 증가)에 해당하므로 연구대상자의 안전을 최대한 보장하여 연구 수행하여 주시기 바랍니다. 연구대상자 수 또는 설문사항 등의 변경이 있을 때에는 반드시 변경신청을 하여 주시고, 연구종료 후 1개월 이내에 종료보고서를 제출하여 주시기 바랍니다. 연구종료 보고 이후 졸업이전에 결과보고(학위논문)를 제출해야 과제관리가 종료됩니다. * 위반/미준수 사례(IRB 승인된 동의서 사용, 연구원 변경, 연구대상자 수 변경, 연구방법 변경, 종료보고 기한 엄수 등)가 발생하지 않도록 연구종료 기간까지 절차를 잘 준수하여 주십시오.					

※ 모든 연구자들은 아래의 사항을 준수하여야 합니다.

1. 승인된 계획서에 따라 연구를 수행하여야 합니다.
2. 위원회의 승인 (IRB 직인이 포함)된 동의서를 사용하여 주시기 바랍니다.
3. 연구의 어떠한 변경이든 위원회의 사전 승인을 받고 수행하여야 하며 연구대상자들의 보호를 위해 취해진 어떠한 응급상황(위해 발생)에서의 변경도 즉각 위원회에 보고하여야 합니다.
4. 위원회의 요구가 있을 때에는 연구의 진행과 관련된 보고를 위원회에 제출하여야 합니다.
5. 위원회가 심의한 과제에 대해 조사 및 감독 차원에서 현장점검을 실시할 시 원활한 점검절차 진행을 위해 연구자는 연구진행과 관련된 서류를 준비하고 협조하여야 합니다.
6. 위원회가 수정 및 보완을 요구한 경우 수정 및 보완 계획을 1개월 이내에 본 위원회에 제출하여야 합니다.
7. 심의결과에 이의가 있을 경우, 심사결과 통지일로부터 2주 이내에 서면으로 이의신청을 할 수 있습니다.
   단, 동일 사안에 대하여 2회 이상의 재심은 하지 않습니다.
8. 총 신청 연구기간이 IRB 연구승인 유효기간을 초과할 경우, 유효기간 만료 2개월 전에 '지속심의' 승인을 받아야 연구지속 진행이 가능합니다.
9. 연구종료 후 1개월 이내에 종료보고를 하여 주시기 바랍니다.
10. 연구와 관련된 기록은 연구가 종료된 시점을 기준으로 최소 3년간 보관하여야 합니다.

연구세대학교 생명윤리위원회 위원장 (인)

ABSTRACT (IN KOREAN)

멀티모달 시스템에서의 사용자 모달리티 선택에 관한 연구

최근 개발되는 많은 기기들은 음성으로 명령을 내릴 수 있는 음성 개인 비서나 음성 기반 대화형 에이전트들을 탑재하고 있습니다. 사용자들은 자신을 목적 달성을 하기위해 기존 사용하던 방식으로 또는 새로운 음성 사용자 인터페이스 (VUI) 중 선택하여 여러 기능을 수행할 수 있습니다. 음성 인터페이스는 눈과 손을 자유롭게 하고 직관적이며, 빠른 속도를 보이는 등 다양한 장점이 있지만, 음성 인터페이스들의 사용량은 점차 감소하고, 사용자들은 적은 영역에서만 음성 상호작용을 사용하고 있다. 이는 기존 반복적인 사용을 통한 경쟁에서 음성 상호작용이 선택되지 못한 것이다. 따라서 기능 및 상황별 적합한 모달리티에 대한 높은 이해를 필요로 하며, 멀티모달 시스템에서 모달리티간 비교를 통해 사용자들이 선택한 모달리티들의 특성을 연구할 필요성이 있다.

본 연구는 멀티모달 시스템에서 사용자의 모달리티 선택을 모달리티의 조작 단위의 조절을 통해 비교하고, 각 모달리티의 특성, 기능의 구조, 그리고 사용 맥락의 영향을 분석 및 평가하는 것을 목적으로 한다. 멀티모달 시스템에서의 모달리티 별 설계 가이드 제시를 위해, 본 논문은 다음과 같은 연구 목표를 달성하고자 한다.

1. 모달리티의 특성과 메뉴 구조가 모달리티 선택과 상호작용 노력 및 만족에 미치는 영향력 분석 및 평가 (세부 연구 1)
2. 멀티모달 시스템 사용 맥락에 따른 모달리티 선택과 상호작용 노력 및 만족의 변화 탐색 (세부 연구 2)

세부 연구 1 의 목표 달성을 위해 다양한 조건의 음성 및 터치 모달리티를 제시하는 멀티모달 시스템을 개발하였고 모달리티 선택 과업을 반복하는 실험이 실시되었다. 터치와 음성 모달리티의 조작 단위를 각각 터치 1 회와 음절 1 개로 정의하였고, 이를 사용하여 터치의 횟수 (1~5 개)와 터치당 음절의 수 (1~8 음절)를 주요 설계 변수로 선정하였다. 또한 과업의 형태를 메뉴 구조 (비계층형, 계층형)라는 주요 설계 변수로 정의하였다. 연구 1 의 실험에서는 주요 설계 변수들의 수준 차이에 따라 사용자들이 선택한 모달리티, 각 모달리티의 상호작용에 필요한 신체적, 정신적 노력, 그리고 해당 상호작용을 통한 과업 수행의 만족도를 평가하였다.

연구 1 의 결과, 사용자의 모달리티 선택은 터치당 음절의 수와 메뉴 구조에 영향을 받는 것으로 분석되었다. 모달리티 선택의 해석을 위한 기준점으로 음성 모달리티 사용량이 50%이상이 되는 지점을 모달리티 스위칭 포인트라 정의하였다. 이러한 모달리티 스위칭 포인트는, 비계층형에서는 터치당 2~3 음절에서, 계층형에서는 터치당 4~5 음절에서 발생하였다. 하지만 터치의 횟수에 따라서는 유의한 차이가 나타나지 않았다. 또한 주요 설계 변수는 모달리티의 신체적, 정신적 노력, 그리고 만족도에 영향을 미쳤으며, 이를 기반으로 사용자의 모달리티 선택이 이루어진 것으로 확인되었다.

세부 연구 2 에서는 멀티모달 시스템의 사용 맥락을 신체적, 정신적 리소스의 사용량에 따라 분류해 총 4 가지 맥락 (기준선, 시청, 독서, 운전)으로 정의하였다. 사용 맥락하에서 모달리티 선택 과업이 진행되었고, 이전 변수와

함께 사용 맥락이 사용자의 모달리티 선택과 주관적 요인(신체적, 정신적 노력, 그리고 만족도)에 영향을 미치는지 확인하였다.

연구 2 의 결과, 사용자의 모달리티 선택은 사용 맥락에 따라 음성을 더 사용하는 방향으로 이동하였다. 모달리티 스위칭 포인트는 기준선 < 시청 < 독서 < 운전의 순서로 더 높은 터치당 음절수로 이동하였다. 즉, 신체적 리소스를 사용하는 맥락이 정신적 리소스를 사용하는 맥락보다 음성 사용량을 더 높였으며, 둘 리소스의 상호작용 효과도 발견되었다. 모달리티의 주관적 요인도 사용 맥락에 의해 증가하였다.

본 연구를 통해 멀티모달 시스템내에서 사용자들의 모달리티 선택은 모달리티간 상호작용 단위의 비율에 의해 결정되며, 이는 기능의 형태나 사용 맥락에 따라 달라짐을 확인하였다.

본 연구에서는 인간공학 및 HCI 분야의 상호작용 단위 기반의 접근방식을 통해 멀티모달 시스템에 대한 평가를 실시하였다. 이를 통해, 멀티모달 시스템에서 사용자들이 음성을 사용하도록 하기위한 음성 모달리티 설계 요인과 지침을 제공하였다. 이러한 결과는 음성 모달리티 특성인 높은 성능, 낮은 인지 부하 등의 장점을 통해 전체 시스템의 사용성을 높이는 설계에 활용할 수 있는 기초 연구로써 의의를 지니고 있다.

핵심어: 멀티모달, 음성 사용자 인터페이스, 모달리티 선택, 상호작용 노력