



Context Matters: Understanding the Effect of Usage Contexts on Users' Modality Selection in Multimodal Systems

Min Chul Cha & Yong Gu Ji

To cite this article: Min Chul Cha & Yong Gu Ji (2023): Context Matters: Understanding the Effect of Usage Contexts on Users' Modality Selection in Multimodal Systems, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2023.2250606](https://doi.org/10.1080/10447318.2023.2250606)

To link to this article: <https://doi.org/10.1080/10447318.2023.2250606>



Published online: 29 Aug 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

Context Matters: Understanding the Effect of Usage Contexts on Users' Modality Selection in Multimodal Systems

Min Chul Cha  and Yong Gu Ji 

Department of Industrial Engineering, Yonsei University, Seoul, Republic of Korea

ABSTRACT

Users often select different modalities in multimodal systems based on context and function. This study aimed to investigate the impact of context on users' modality selection in a multimodal system. A modality selection task was presented through the developed multimodal system, where the task's difficulty was controlled based on the number of syllables and touches. The experiments were examined in four usage contexts (baseline, watching, reading, and driving), and the selected modality, physical and mental effort, and satisfaction were measured. A regression model for predicting modality selection was established using binomial logistic regression. The results showed that voice usage increased in the order of baseline, watching, reading, and driving contexts. It was revealed that these changes were attributable to differences in physical and mental interaction efforts of modalities according to contexts. Our finding provides valuable insights into users' modality selection in different contexts, which could inform the design of more efficient multimodal systems.

KEYWORDS

Multimodal; modality selection; usage context; touch interaction; voice modality

1. Introduction

Recently, many devices have incorporated multimodal systems that allow users to operate them through both voice and touch. When using these systems, users select the preferred modality to execute the desired function, such as playing music by tapping their smartphone or using voice commands like "Hey Google, play songs by Ariana Grande." While each modality has unique advantages and disadvantages, users can benefit from both by selecting the modality that is most appropriate for their situation. Touch input provides a familiar and immediate operation, while voice input enables hands- and eyes-free operation. Depending on their task, purpose, and context, users switch between modalities to achieve their goals. For example, people have two options: get up and flip the light switch, or say "Alexa, turn off the lights". Normally, they would flip the physical switch, but before going to bed, they would ask Alexa to turn off the lights.

Overall, the design of successful and usable multimodal systems remains a challenge for human-computer interaction (HCI) researchers, as user's modality selection or preference can vary widely across different contexts and usage scenarios. For example, Lemmelä et al. (2008) created a multimodality application prototype that provided both voice and touch modalities for text messaging functions in mobile contexts to investigate which modalities users use in walking and driving contexts. In their research, voice interaction was preferred for driving context, but not for walking. Reicherts et al. (2022) explored the effect of voice and touch modalities in which the agent interacts with users on the user's behavior. They found that a voice interface encouraged more interactions with an

agent and more data visualizations explored than a touch interface. Hoffmann et al. (2019) proposed input modalities, including voice, mid-air gestures, and touch, for smart homes, and users evaluated the fitness of modalities differently depending on the tasks.

Research on finding the suitable modality in various situations is being actively conducted, but existing studies have focused on the results caused by each modality. They have claimed that, depending on the situation, users should use a specific modality, by focusing on differences in modality-specific task performance measures such as total task time, success rate (Beckers et al., 2014; Lee & Lai, 2005), subjective measures such as usability, workload, preference (Baxter et al., 2021; Detjen et al., 2020; Schartmüller & Riener, 2022; Zhao et al., 2021), or the effects that other tasks receive depending on the modality, such as trust for system, detection response task reaction time, lateral lane position (Qiu & Benbasat, 2005; Tsimhoni et al., 2004). While these results may be appropriate for appealing to the advantages of each modality, they are not sufficient for answering why users chose a specific modality in their usage context.

This study aims to investigate the effects of usage contexts on the user's modality selection in a multimodal system. By deepening our understanding of users' multimodal interaction and usage context, we seek to help designers develop multimodal systems that are based on people's natural behaviors. To achieve this goal, we developed a multimodal system that provides both voice and touch input modalities and explored changes in user modality selection, physical and mental effort, and satisfaction across varied contexts of daily life. Additionally, we proposed an interaction unit-based approach

to generate tasks for testing a multimodal system and analyze the reasons why users choose each modality. The results of this study are expected to aid in the interpretation of users' modality selection and in designing systems that encourage the use of specific modalities.

2. Related work

2.1. Modality selection between voice and touch

Users select the modality for a task and usage context based on their repetitive task performance experiences in multimodal systems (Suhm et al., 1999), and the efficiency and effectiveness of the modality influence their modality choices. From the perspective of input modality, effectiveness is related to user input accuracy (Card et al., 1990; Chen & Tremaine, 2006), and is a factor focused on technical performance. In terms of efficiency, a modality is considered efficient when the speed at which commands are entered is fast or a task can be performed with less effort (Perakakis & Potamianos, 2008).

This study focused on how modality selection changes based on the efficiency level of each modality, specifically targeting voice and touch, which are commonly used input modalities in various devices and applications. To compare efficiency between modalities, this study proposed an approach based on interaction units of voice and touch input modalities. The unit of interaction for touch input is relatively clear, since it is defined by the number of touches. In the traditional approach, the keystroke level modeling (KLM) method (Card et al., 1980), one keystroke on the keyboard was defined as one operating unit. Lee et al. (2019) expanded this to define one tapping as an operator unit and predicted the task time of the touch interaction.

On the other hand, there is limited research from the HCI perspective on the interaction unit of voice input. While some studies have investigated subjective or objective results induced by voice interaction, they have only presented text entry or their system-specific tasks (Cherubini et al., 2009; Hauptmann & Rudnick, 1990; Wu et al., 2015). However, as noted in the example of music playback, input attributes can differ when voice and touch perform the same function in an actual system (Zhao et al., 2021). Therefore, the operation units of voice and touch must be considered individually. A commonly used voice unit is a word, and the speed of input with other modalities is often compared through uttered words per minute (WPM) (Foley et al., 2020; Ruan et al., 2018). Schaffer et al. (2011) conducted a study on modality selection and mental effort using voice and touch interfaces. They found that the usage of the voice interface increased as the interaction step of the touch interface increased, but the interaction unit of the voice interface was treated as a single step.

There are several units of language such as letters, syllables, and words, and all of them can be used as units for designing a voice interface. However, among existing human language studies, Malaia and Wilbur (2020) suggested syllables as a unit of information transfer for humans in linguistic communication. They insisted that humans subdivide words into syllables in the process of recognizing or uttering

them based on neuroscience. As such, in terms of speech synthesis or speech recognition (S/R), the unit of speech generation is a syllable, and it has been studied that the composition of these syllables is produced through specific rules depending on the language (Fujimura, 1975; Sendlneier, 1995). In the case of English, multiple letters within one word combine to form a syllable, while in Korean, as shown in Figure 1, a character is composed of multiple phonemes and one such character serves as a syllable (Simpson & Kang, 2004). Ferrand (2000) investigated the effect of the number of syllables in the naming of French words. He found that the number of syllables affected both immediate and delayed utterances, and he suggested the syllable as a unit of processing in the recognition of these stimuli. Therefore, in this study, such syllables are defined as the interaction unit of voice input modality, and since this is a study based on Korean, one character in Korean served as one syllable.

2.2. Multimodal usage context

When using a multimode system, users tend to select the most efficient mode depending on the situation, and this is an essential element of interface design (Kocaballi et al., 2019). Additionally, users are often required to multitask in various contexts, which can affect their behaviors, task performance, and workload (Kim et al., 2020; Mustonen et al., 2004). To explain the effects of context and multitasking, researchers commonly used the human information process perspective. One representative model is Wickens' multiple resource model (Wickens, 2008), which considered interference caused by resource overlap between tasks in multitasking situations. The model suggested that resource demand and conflict can lead to overload conditions for users. Wickens also found that cross-modal multitasking enables parallel information processing through time-sharing, while inter-modal multitasking can cause interference due to the sequential processing of modalities.

Previous studies have also taken a resource-based approach to analyze the context or task and its effect on user behavior. For example, Yoon et al. (2021) studied the effect of non-driving related tasks, such as video watching, reading, texting, and cellphone talk, on the take-over time of autonomous vehicles by analyzing their physical and mental resource demands in driving situations. Lemmelä (2008) proposed an approach to finding optimal modalities in mobile and pervasive environments. They stated that multimodality can support user in different context by using perceptual channels, which are currently less occupied than

| | | |
|------------------|--------|-------|
| English Word | school | |
| Korean Word | 학교 | |
| Korean Syllables | 학 | 교 |
| Korean Phonemes | ㅎ ㅏ ㄱ | ㄱ ㅛ |
| Pronunciation | h a k | k y o |

Figure 1. An example of the composition of words and syllables of the Korean Hangul.

others. Later, Lemmelä classified usage contexts by a person's aural, visual, physical, and cognitive load and showed that the context affects users' modality preferences (Lemmelä et al., 2008). Jameson and Klöckner (2005) analyzed the multitasking context by considering humans' resources of hands, eyes, voice, working memory, and ears and focused on the resource conflict between system-related and environment-related tasks due to limited physical and cognitive resources. Therefore, this study expected modality selection to vary according to the user's context in a multimodal system and analyzed its cause.

2.3. Importance of interaction efforts and satisfaction in modality selection

The modality selection is influenced by the effort and workload required for interaction with the system through each modality (Huang et al., 2021; Liu & Thomas, 2017). Budi (2013) defined the interaction cost as the total physical and mental effort necessary for interaction, and users tend to minimize it by switching modalities for maximizing the effectiveness of their actions. Jeon et al. (2015) evaluated the effectiveness of auditory menu cues for menu navigation in vehicles, and while there was no difference in primary task performance, users preferred auditory cues with lower workload scores. In contrast, in the study by Kim et al. (2019) on error recovery strategies for voice user interfaces, the strategy requiring the highest workload was the most preferred.

In the field of HCI, various measures have been developed to measure the efforts of interaction. One of the most popular measures is the Nasa-TLX, which measures task workload in seven dimensions (Hart & Staveland, 1988). Each dimension is measured by a single questionnaire and can be used in combination or individually (Arrabito et al.,

2015; Hwangbo et al., 2013; Laureiti et al., 2017; Turner et al., 2021). Examples of such methods that measure effort in a single questionnaire include usability magnitude estimation (UME) or subjective mental effort question (SMEQ) (McGee, 2004; Zijlstra, 1993). Sauro and Dumas (2009) found that the SMEQ performs better than UME, and in this study, SMEQ form, which is more intuitive to evaluate than Nasa-TLX because subjects had to repeat the subjective questionnaire several times, was selected.

The study also assumed that the difference in efficiency between these modalities would affect satisfaction. According to ISO 9241-11 (ISO, 1998), satisfaction is one of usability factors along with effectiveness and efficiency. Calisir and Calisir (2004) found that several characteristics of interfaces affect satisfaction and measured satisfaction on a five-point scale. Similarly, in existing studies, satisfaction is often measured by a single survey question with a five- or seven-point Likert scale or a 0–100 scale (from very dissatisfied to very satisfied) (Findlater & McGrenere, 2008; Park et al., 2019). Therefore, in this study, we measured satisfaction in the same way to see if there is a difference in satisfaction between modalities.

3. Methodology

A laboratory experiment was conducted using a newly developed multimodal system to investigate the impact of usage context on users' modality selection. We hypothesized that the context of use would influence the interaction effort and satisfaction required for users to interact with multimodal systems, ultimately affecting their modality selection. Figure 2 presents the conceptual model for this study, and this section provides detailed information on the experimental procedures.

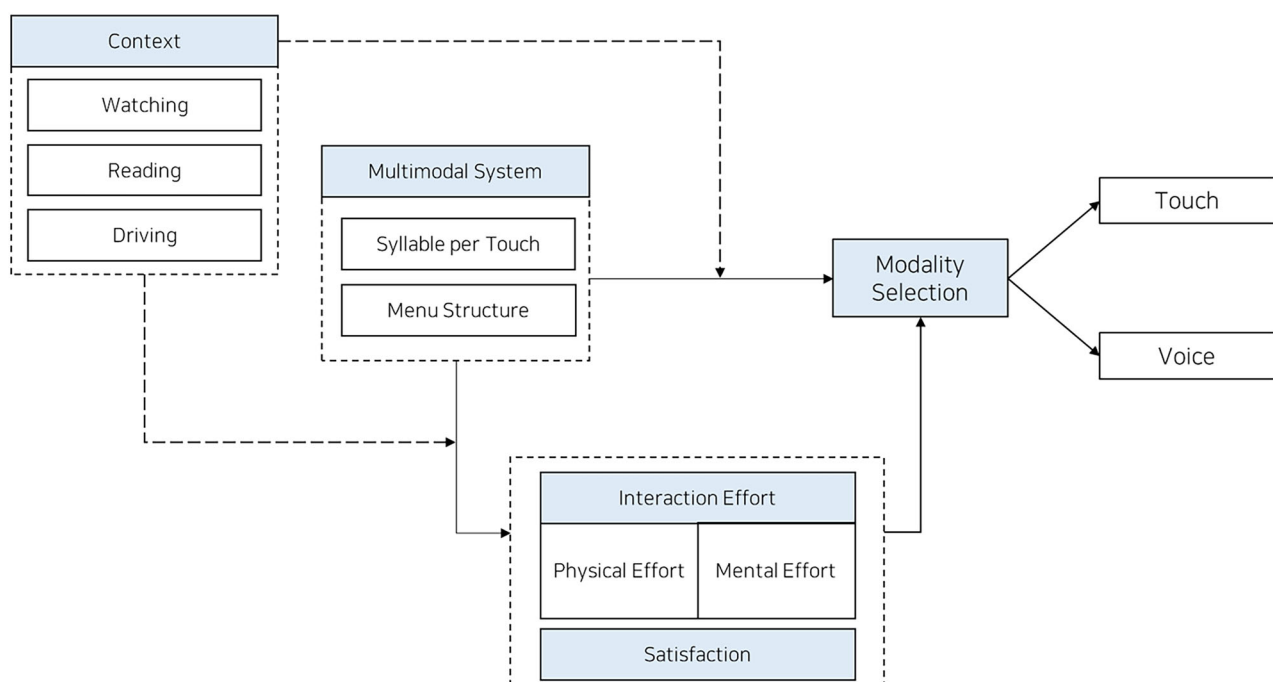


Figure 2. Conceptual model of the current study.

3.1. Participants

Thirty-one participants (17 males and 14 females; mean age = 27.3, $SD = 4.19$) from the local university participated in the study. All participants had used a voice interface or voice service before, with an average of 1.8 platforms (Google, Alexa, Bixby, etc.) used. However, due to technical issues, the baseline data of one participant were not recorded, and one participant could not perform a reading condition. Excluding them, a total of 3043 data points in the non-hierarchical condition and 4878 data points in the hierarchical condition were used in the analysis. The experimental procedures were approved by the university's Institutional Review Board (IRB no. 7001988-202210-HR-1709-02) and were explained to each participant before beginning the experiment. All participants signed informed consent and received monetary compensation for their participation.

3.2. Contexts

Based on the literature review, we selected three contexts (reading, watching, and driving) commonly used in existing research. For the experiment, human physical and mental resource demands of each context were categorized, and a baseline context was added. Finally, four contexts were presented to participants (as shown in Figure 3). A detailed explanation of each context follows:

1. *Baseline*: Participants performed only the modality selection task without additional context.

2. *Reading*: Participants selected a book and read it while performing a modality selection task using a touch screen. They were instructed to hold the book with both hands in front of them and read at their own pace. Additionally, they were asked to position the book at the center of their body and perform the modality selection task on the right side of the touchscreen using only their right hand.
3. *Watching*: Participants were asked to watch an animated film titled "Sing" on a tablet PC while performing a modality selection task. The movie continued to play during the task, and participants were free to move their hands but instructed to use only their right hand for the modality selection on the right touchscreen.
4. *Driving*: For the driving context, City Car Driving software (Forward Development) was employed. A quiet rural highway drive with a total of six lanes, three 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/hr. They held the steering wheel with both hands while driving and used only their right hand to perform the modality selection task as in other contexts.

3.3. Apparatus

A multimodal system, capable of both voice and touch interactions, was developed using JavaScript. The program was presented to participants through a laptop and a portable monitor with a 15.6-inch FHD touch screen (see Figure 4(a)).



| Contexts |  |  |  |  |
|----------|---|---|--|---|
| | Baseline | Watching | Reading | Driving |
| Physical | - | - | Postural | Postural, Locomotion |
| Mental | - | Visual, Auditory | Visual | Visual, Auditory |

Figure 3. Four experimental contexts and the physical and mental resource demands of each context.



Figure 4. (a) Experimental setup and (b) example of experiment in driving context.

Participants completed both the modality selection task and the questionnaire using it.

Additional equipment was used to implement the experimental contexts. For reading, five books composed of novels and essays were provided. For watching, a Samsung Galaxy Tab S7 Plus was used. For driving, a Logitech G27 steering wheel and a 27-inch LG monitor were installed. The driving context was implemented using City Car Driving software as a driving simulator to create a rural highway driving environment.

3.4. Task design

3.4.1. Speech material

To prevent the use of unfamiliar vocabulary or unnatural word placement, we used restaurant names registered in the local city as stimuli for the voice modality. A refining process was performed on approximately 480,000 restaurants to remove special characters and numbers. Restaurants with a maximum length of 20 characters were used as stimuli, resulting in a list of 430,276 restaurants. If an experimental condition required more than 20 characters of speech material, two restaurants were combined to generate the length of syllables required for the experiment.

3.4.2. Modality selection task

A modality selection task consisted of repeating a series of processes in which subjects performed a given trial by selecting a modality between voice and touch. There were five trials per condition, and in each trial, they were asked to either say the name of a given restaurant or touch buttons. The menu structure for the task, as shown in Figure 5, consisted of both non-hierarchical and hierarchical menus, with touch modality buttons located at the top and S/R buttons for voice modality located at the bottom. The size of each button is designed at a level that does not affect touch performance, with a square size of 30 mm × 30 mm button

spacing of 3 mm (Hwangbo et al., 2013). The description of each modality is as follows.

3.4.2.1. Touch modality. In our previous research, the number of touches was found to have no effect on modality selection, so the number of touches was fixed at three. The touch modality tasks are designed differently depending on the two menu structures. In the non-hierarchical menu (Figure 5(a)), the touch modality imitates tasks involving repetitive manipulation after cognitive judgment, such as adjusting volume, temperature, and radio channels. A random number between 10 and 90 was presented as the “Target Number”, and the participant adjusted the “Present Number” to match the “Target Number” by touching either the + or - button. In the hierarchical menu (Figure 5(b)), the touch modality presented a task that required users to search for new items based on the hierarchy whenever they touch. Nine items (phone call, weather, music, order, brightness, air conditioner, radio, volume, and navigation) are randomly placed on a 3 × 3 grid, and one of them was presented as the touch target. When participants touch the target, the grid was randomly rearranged, and they needed to search for the touch target again.

3.4.2.2. Voice modality. When participants decided to use voice modality in the modality selection task, they initiated S/R by touching the “S/R” button and they said the name of the restaurant displayed on the screen. A “beep” sound signals the start of S/R. The S/R system was implemented using the Google STT API, with a 2.5% chance of error. The length of targets was adjusted to have 1–8 syllables per touch (S/T), so the presented restaurant names had 3, 6, 9, 12, 15, 18, 21, and 24 characters.

3.5. Experimental design and procedure

All experimental conditions in this study were conducted using a within-subject design. Two menu structures were

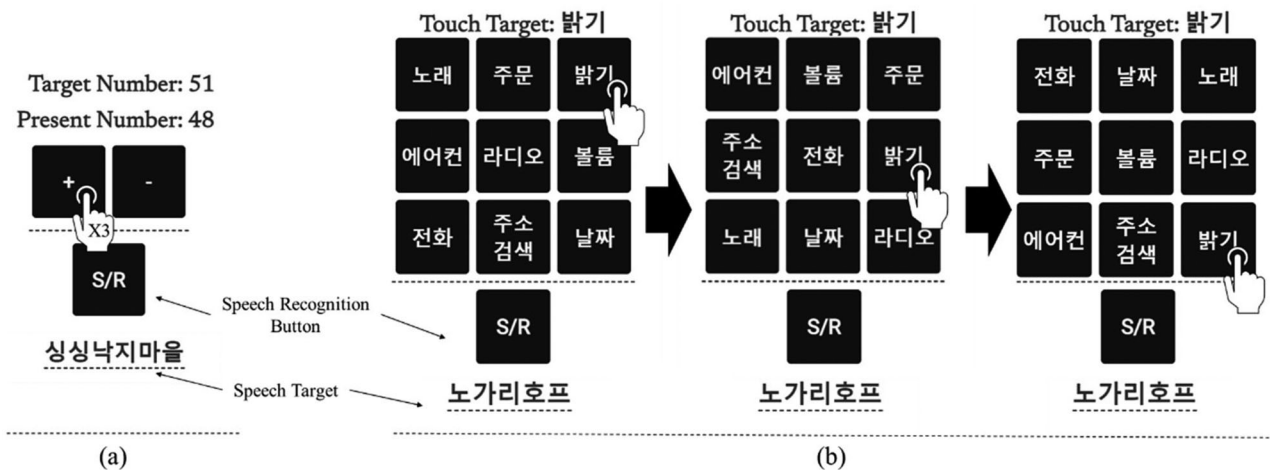


Figure 5. Task examples of (a) non-hierarchy and (b) hierarchy conditions.

used: non-hierarchical and hierarchical. The length of voice targets was manipulated, ranging from 1 to 8 S/T depending on the menu structure (non-hierarchical: 1–5 S/Ts; hierarchical: 1–8 S/Ts). Since the number of target touches was fixed at three times, restaurant names were presented as voice targets with 3, 6, 9, 12, 15, 18, 21, and 24 syllables, which corresponds to the number of characters in Korean. Each participant performed five trials of the modality selection task for each S/T condition. Additionally, the study included four usage contexts: baseline, reading, watching, and driving. Consequently, participants performed 260 modality selection tasks (4 contexts \times {5 S/Ts (non-hierarchy) + 8 S/Ts (hierarchy)} \times 5 trials = 260 trials) per participant.

Figure 6 illustrates the procedure used in the current study. The experiment began with obtaining consent from the participants and providing them with detailed explanations about the experiment. Demographic information was collected before proceeding to the main experiment. The baseline context was performed to train the participants on the modality selection task, and then the remaining three contexts (reading, watching, and driving) were randomly presented. Within each context, the non-hierarchy condition was presented first among two menu structures, and the order of S/Ts was randomly set. The modality selection task was performed five times in each condition, and the participant had to choose the modality considered to be more efficient to proceed to the next step, either voice or touch modalities. In each trial, the target number or panel arrangement changed for the touch modality, and the target restaurant changed for the voice modality. Participants were instructed to perform the modality selection task whenever possible during the context task. After completing all the modality selection tasks, the physical effort, mental effort, and satisfaction of each modality required to perform the corresponding S/T conditions were evaluated. For the voice modality, a subjective questionnaire was conducted after

each S/T condition, while for the touch modality, it was randomly measured in only half of the conditions.

3.6. Measures and analysis

Four dependent variables were measured in the study, including one objective variable and three subjective variables. The objective variable was the modality chosen by the participant in the task, which was recorded as either voice or touch. To evaluate the physical and mental effort required to use each modality, two modified questionnaire items based on the SMEQ were used (in Appendix A). The questionnaires used a 16-point scale ranging from 0 to 150, and participants were asked to evaluate how much physical and mental effort it took to use each modality. They were instructed to exclude ratings of the PTT button when evaluating voice modality. The questionnaires assessed both the physical resources required (such as moving a hand or making a voice) and the mental resources required (such as making a decision or calculating numbers). Additionally, the satisfaction of interacting with each modality was measured. The satisfaction question used in Park et al. (2019) was adapted to "How satisfying was it to perform the tasks in each modality?" with a scale of 0–100.

Two hundred and sixty modality selection data points per subject were collected, and the effect of usage context on modality selection was analyzed using binomial logistic regression. A prediction model was established to predict the rate of voice usage, where voice modality was coded as "1" and touch modality was coded as "0". The usage context was analyzed as a categorical variable, with the baseline context used as a reference category.

A three-way ANOVA was performed to examine the effects of context on physical and mental interaction efforts and satisfaction. The interaction efforts and satisfaction of each modality and context were averaged for the analysis.

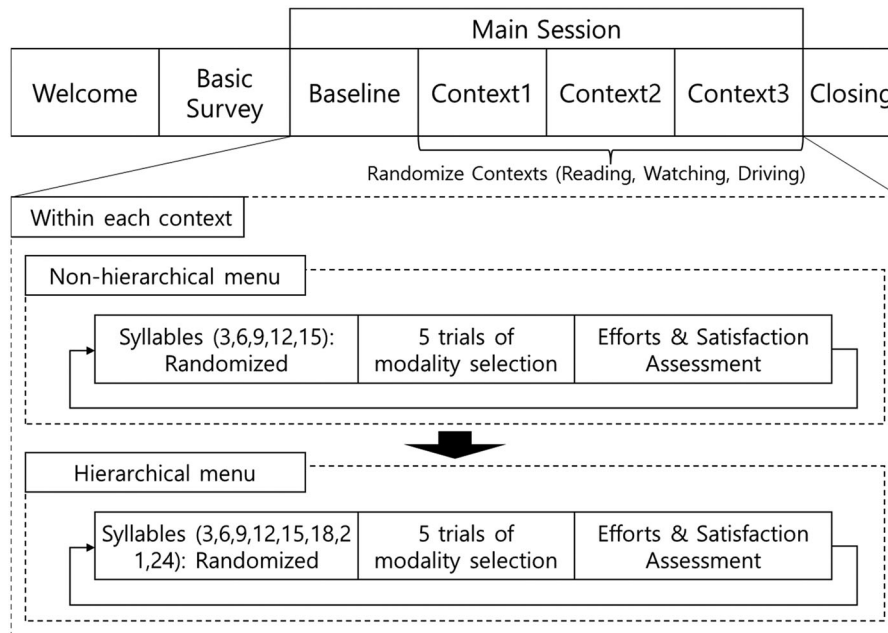


Figure 6. Overall procedure of the current study.

The statistical significance level for all tests was set to 0.05. The statistical analysis was conducted using IBM SPSS Statistics version 26.0 (Armonk, NY).

4. Results

4.1. Non-hierarchical menu

The study revealed that in the non-hierarchical menu structure, participants' modality selection shifted from voice to touch as the S/T increased. We found that there is a point where the usage ratio of voice and touch modalities crosses, that is, a point where each modality usage becomes 50%. We call this point the "modality switching point", where the dominant modality switches. To further examine the differences in the modality switching point across different contexts, a detailed binomial regression analysis was performed.

The binomial logistic regression with forward selection (LR) was conducted to analyze the relationship between the number of syllables, user context, and modality usage, as shown in Table 1. The results indicated that both syllables and all contexts were significant variables predicting the rate of voice usage at the significance level of 0.05. Specifically, it was found that each additional increase of one S/T (equivalent to an increase of three syllables) was associated with a 56% decrease in the odds of using voice (OR: 0.440; 95% CI:

0.411–0.471). Furthermore, the study found that when driving, reading, or watching, people used the voice modality 5.381 times (95% CI: 4.189–6.912), 2.231 times (95% CI: 1.761–2.828), and 1.39 times (95% CI: 1.103–1.754) more than the baseline, respectively.

Figure 7 displays the voice usage predicted by the logistic regression model in non-hierarchy. The modality switching point was indicated by the 50% auxiliary line. The results showed that the modality switching point was approximately 2.5 S/T in baseline, three S/T in watching, nearly four S/T in reading, and nearly five S/T in driving. The model's classification accuracy was 74.3%. The ROC curve's AUC for the model was 0.801, indicating an excellent ability to discriminate between the classes.

4.2. Hierarchical menu

The modality selection in the hierarchical menu condition followed a similar trend to that in the non-hierarchical menu condition. The results of logistic regression analysis for the hierarchical menu conditions are summarized in Table 2. Syllable per touch and all contexts were significant predictors of voice modality use. Holding contexts constant, the odds of using voice modality decreased by 55.8% (95% CI: 0.423–0.462) for each additional increase of one S/T (equivalent to an increase of three syllables). When driving

Table 1. Results of binomial logistic regression in the non-hierarchy condition.

| IV | | B | SE | Wald | p | OR | 95% CI | |
|----------------------|----------|--------|-------|---------|----------|-------|--------|-------|
| | | | | | | | Lower | Upper |
| Syllables 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 |
| | | | | | | | | |

** $p < 0.01$.

*** $p < 0.001$.

^aReference group: context × baseline.

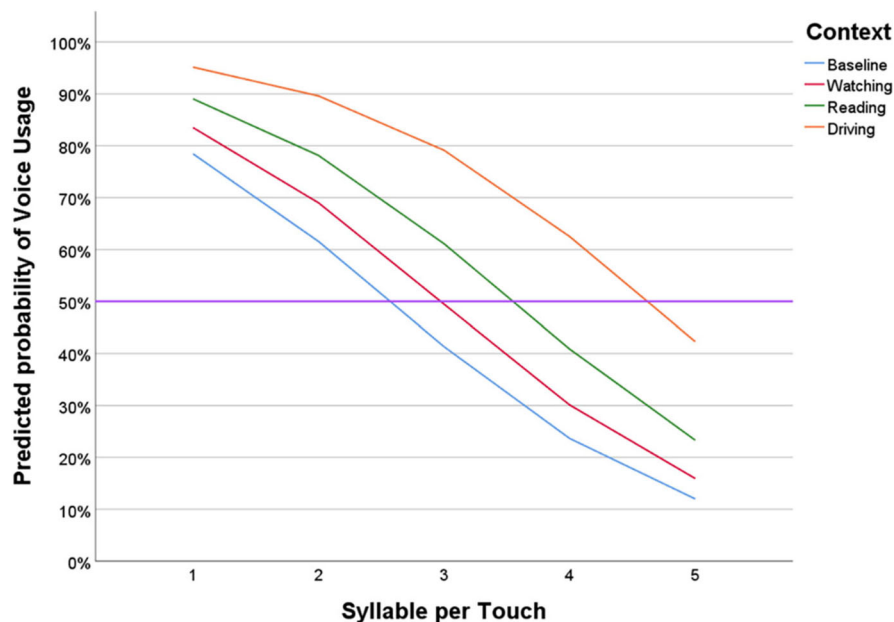
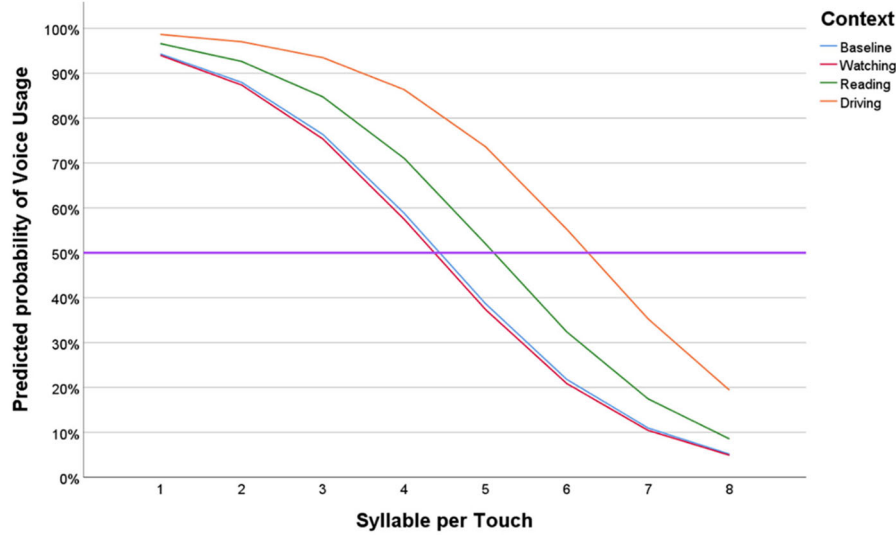


Figure 7. Predicted voice usage according to contexts in non-hierarchy.

Table 2. Results of binomial logistic regression in the hierarchy condition.

| IV | | <i>B</i> | SE | Wald | <i>p</i> | OR | 95% CI | |
|----------------------|----------|----------|-------|----------|----------|-------|--------|-------|
| | | | | | | | Lower | Upper |
| Syllables 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 |

*** $p < 0.001$.^aReference group: context \times baseline.**Figure 8.** Predicted voice usage according to contexts in hierarchy.

or reading, participants used voice modality 4.431 times (95% CI: 3.547–5.535) and 1.719 times (95% CI: 1.388–2.128) more than the baseline, respectively. However, in contrast to the non-hierarchical menu condition, there was no significant difference between baseline and watching a video.

The predicted voice usage in hierarchy menu condition is shown in Figure 8. In the hierarchical menu condition, the modality switching point shifted to the right compared to the non-hierarchical menu condition. Modality switching occurred at about 4.5 S/T for baseline and watching, five S/T for reading, and 6.5 S/T for driving. The model classified the selected modality with an accuracy of 79.3%, and the AUC of ROC curve was 0.879, indicating excellent discriminating ability.

4.3. Interaction efforts and satisfaction

4.3.1. Physical interaction effort

Figure 9 shows the physical effort of voice and touch modalities. The physical effort of touch (blue bar) increased more rapidly than those of voice in each condition according to contexts.

A three-way ANOVA was conducted to analyze the effects of menu structure and context on the physical interaction effort of each modality. The results of the ANOVA are presented in Table 3 and revealed significant main effects of menu structure and context ($p_{(menu\ structure)} = 0.001$, $p_{(context)} < 0.000$). Additionally, among the

interaction effects, only the interaction effect of context and modality was significant ($p = 0.005$).

To further analyze the main effect of both variables in each menu structure, simple main effect analysis and post-analysis were conducted due to the significant interaction effect between context and modality. Both menu structure conditions exhibited a significant main effect of context of use on the physical effort required for both voice and touch modalities (all $ps < 0.000$). Bonferroni's pairwise comparison revealed a significant difference in physical effort between the voice and touch modalities in the driving context for both the non-hierarchical and hierarchical menu structures ($p_{(non-hierarchy, driving)} = 0.010$, $p_{(hierarchy, driving)} = 0.037$). The graph in Figure 10 indicates that the average physical effort of touch increased more rapidly than that of voice and was more affected by contexts.

4.3.2. Mental interaction effort

Figure 11 illustrates the mental effort of voice and touch modalities for each experimental condition. The increase in mental effort for voice and touch modalities tends to follow a similar pattern to that observed in physical effort.

The results of the ANOVA on mental efforts are summarized in Table 4, which indicates that both menu structure and context had significant main effects ($p_{(menu\ structure)} < 0.000$, $p_{(context)} < 0.000$). Only the interaction effect between context and modality was significant among the interaction effects ($p = 0.014$).

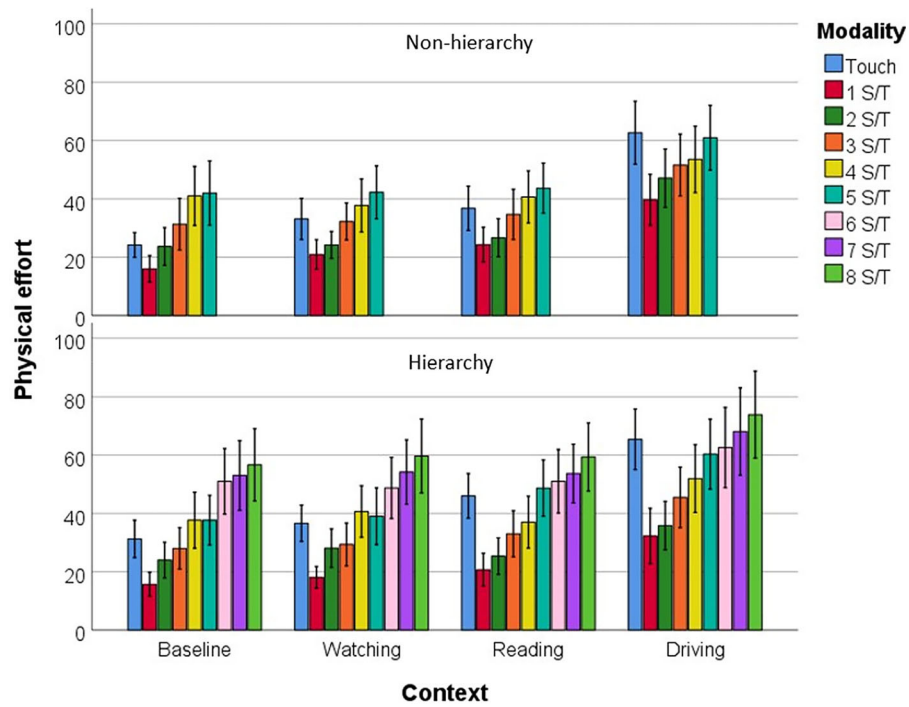


Figure 9. Physical effort results for modality, menu structure, and context.

Table 3. 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 \times context | 3 | 0.358 | 0.783 | 0.001 |
| | Menu structure \times modality | 1 | 0.046 | 0.830 | 0.000 |
| | Context \times modality | 3 | 4.339 | 0.005** | 0.007 |
| | Menu structure \times context \times modality | 3 | 0.147 | 0.932 | 0.000 |

** $p < 0.01$.

*** $p < 0.001$.

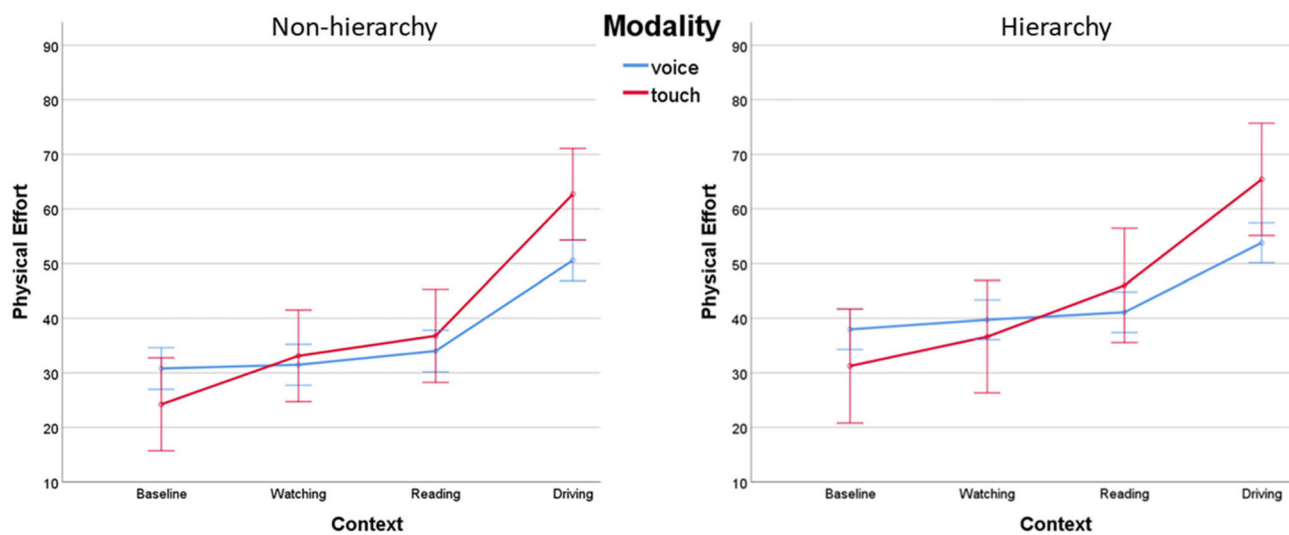


Figure 10. Average physical efforts of each modality by contexts.

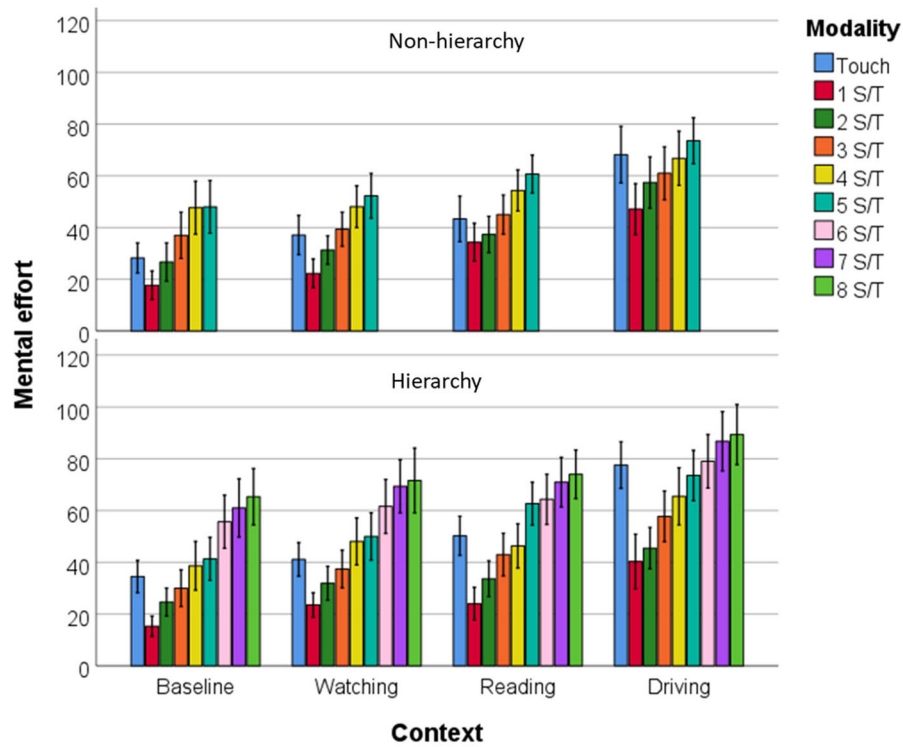


Figure 11. Mental effort results for modality, menu structure, and context.

Table 4. Result of three-way ANOVA for mental effort by each condition.

| 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 \times context | 3 | 0.036 | 0.991 | 0.000 |
| | Menu structure \times modality | 1 | 0.020 | 0.888 | 0.000 |
| | Context \times modality | 3 | 3.531 | 0.014* | 0.006 |
| | Menu structure \times context \times modality | 3 | 0.322 | 0.809 | 0.001 |

* $p < 0.05$.

*** $p < 0.001$.

The average mental effort for each modality and menu structure according to the context is shown in Figure 12. Given that the ANOVA results for mental effort were comparable to those for physical effort, a similar approach was used to analyze the main effect of context and modality. Simple main effect and post hoc analyses were conducted to examine the main effects of both variables.

A simple main effect of context was significant in both modalities (all $ps < 0.000$). However, in the Bonferroni pairwise comparison, there was a marginal difference in mental effort between modalities in the driving context of the hierarchical menu structure ($p_{(hierarchy, driving)} = 0.056$). Mental effort increased from baseline to watching, reading, and driving. Although the difference between both modalities was not statistically significant, there was a crossing point between the mental efforts of voice and touch (see in Figure 12).

4.3.3. Satisfaction

The average satisfaction for each experimental condition is presented in Figure 13. Interestingly, the results of

satisfaction differed from those of physical and mental effort. Specifically, in the voice modality, the satisfaction scores were consistent across both S/T conditions and contexts. However, for touch modality, satisfaction decreased according to context.

Three-way ANOVA with satisfaction was also performed for detail analysis. The results of ANOVA are summarized in Table 5. Unlike interaction efforts, there was no main effect of menu structure, but the main effects of context and modality were significant context ($p_{(context)} = 0.015$, $p_{(modality)} = 0.012$). Only the interaction effect between context and modality was significant among the interaction effects ($p < 0.000$).

The average satisfaction according to the context and modality of each structure is shown in Figure 14. To analyze the main effect of context and modality on satisfaction, simple main effect and post hoc analyses were conducted.

The simple main effect of context on satisfaction of touch was significant in non-hierarchy and marginal significant in hierarchy ($p_{(non-hierarchy)} < 0.026$, $p_{(hierarchy)} < 0.059$). Interestingly, the satisfaction for voice modality remained

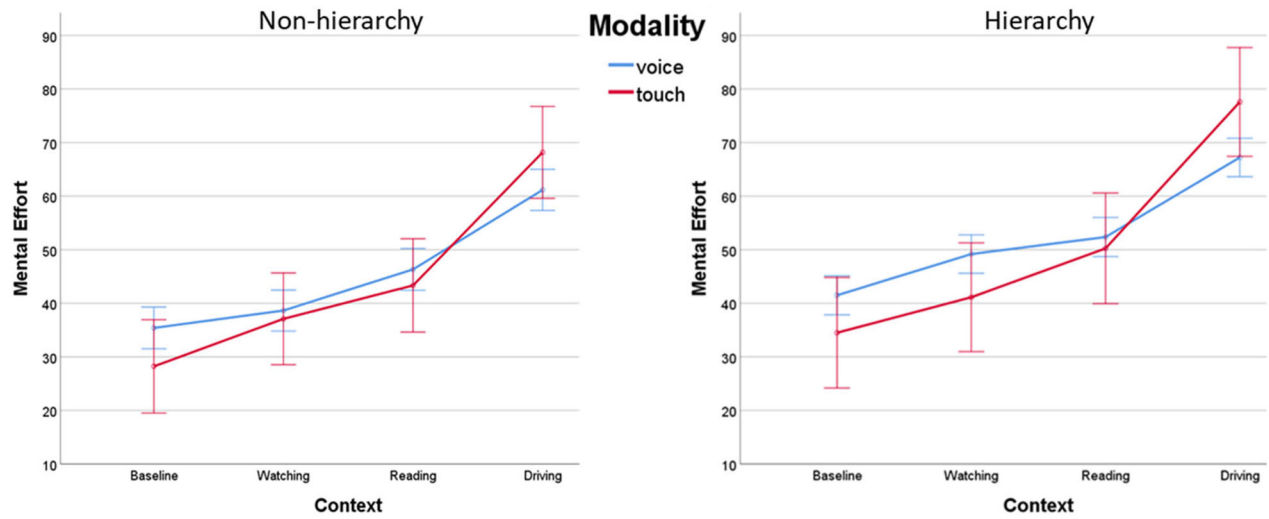


Figure 12. Average mental efforts of each modality by contexts.

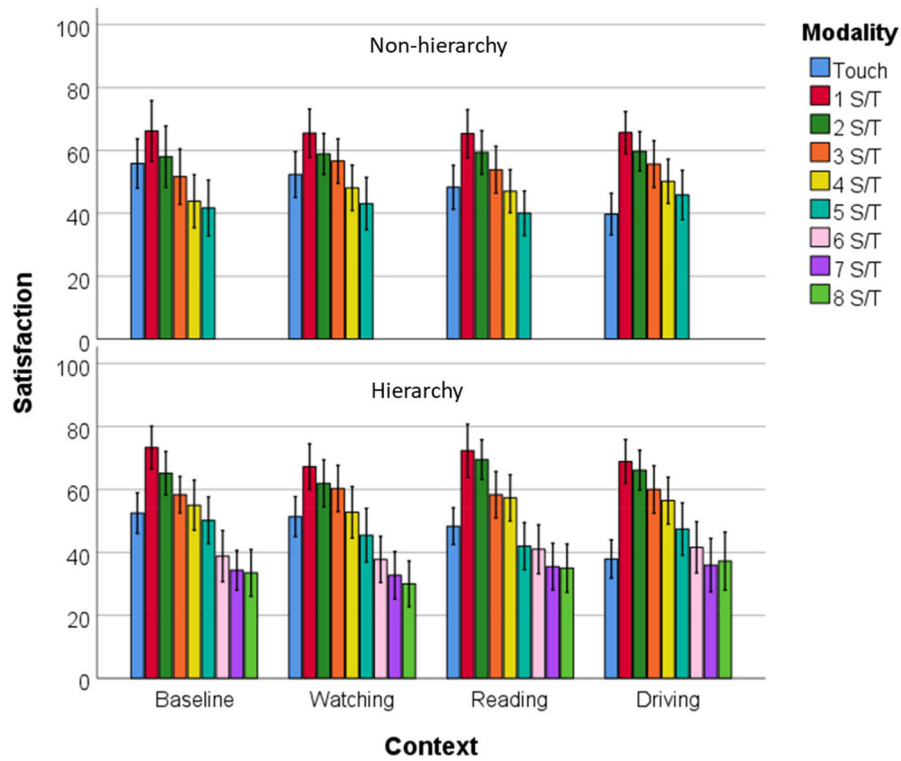


Figure 13. Satisfaction results for modality, menu structure, and context.

Table 5. Result of three-way ANOVA for satisfaction by each condition.

| 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 \times context | 3 | 0.121 | 0.948 | 0.000 |
| | Menu structure \times modality | 1 | 0.261 | 0.609 | 0.000 |
| | Context \times modality | 3 | 6.007 | 0.000*** | 0.010 |
| | Menu structure \times context \times modality | 3 | 0.214 | 0.887 | 0.000 |

* $p < 0.05$.

*** $p < 0.001$.

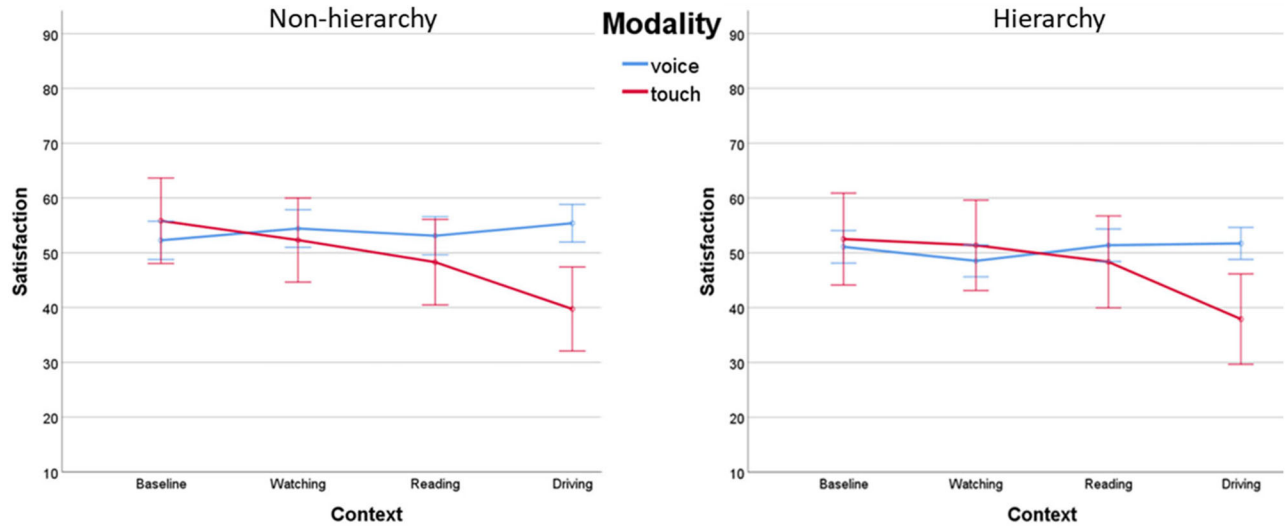


Figure 14. Average satisfaction of each modality by contexts.

constant across all contexts and significantly differed from touch in driving ($p_{(\text{non-hierarchy, driving})} < 0.000$, $p_{(\text{hierarchy, driving})} = 0.002$).

5. Discussion

This study investigated how usage context influences users' modality selection in four scenarios: baseline, watching, reading, and driving. Through experimentation, we found a point where the user's selection rates of the two modalities intersect depending on the condition. We called that point the "modality switching point" and it shifted according to S/T, menu structure, and context. When S/T was smaller than the modality switching point, the user's voice modality usage became dominant. Therefore, this point can be used as a reference point to determine the dominant modality in various conditions. This study suggests that even in the same multimodal system, the modality switching point could change based on the user's context. These findings emphasize the importance of the usage context when designing multimodal systems.

In this study, the interaction unit of voice and touch was defined as one syllable and one touch, respectively. Using this approach, the modality that users would use to perform tasks could be accurately predicted based on the ratio of syllables and touches. In the baseline condition, where no additional context was given, the modality switching points were approximately 2.5 S/T in the non-hierarchical menu and about 4.5 S/T in the hierarchical menu. It was found that voice modality was predominant in cases where syllable per touch was lower than those modality switching points. This study built upon the efficiency perspective approach to modality usage presented in previous studies and resolved the issue of ambiguous voice modality operational units (Perakakis & Potamianos, 2008; Wechsung et al., 2010). The predictive model of this study can be used to predict the modality that users will select based on the minimum length of speech or touch needed to execute a function. Furthermore, it also can help to interpret the variations in

voice usage across various functions due to modality constraints. The study also raises questions about natural speech, which has been the goal of voice interface design in previous studies (Hua & Ng, 2010; Kim et al., 2021). While many devices encourage natural speech, such as asking "Play the news. What's the latest news from BBC?", research suggests that such long voice commands can actually reduce voice usage. Therefore, encouraging short utterances like "Latest news from BBC" may be a way to promote the use of voice interfaces.

This study found intriguing findings regarding the user's modality selection based on usage contexts. Figures 7 and 8 demonstrate that the modality switching point shifted to the right in the presence of context. This suggests that even when the number of S/T increases, i.e., the user speaks more, users still prefer voice over touch in a multitasking context. These findings are consistent with previous research on the comparison of multimodality in walking and driving (Lemmelä et al., 2008) and provide additional insight into the preference shift in modality based on context. Most multimodal systems or devices provide the same modality selection option for interface consistency and uniformity to users, regardless of context. However, this study suggested that even when performing the same function in the same multimodal system, users may choose different modalities based on context. This change in user behavior could impact the modality satisfaction and ultimately the usability of the system.

The results of Section 4.3 showed that the change in modality selection based on usage context was due to the impact of usage contexts on the physical and mental interaction effort required for each modality. There was a significant interaction effect between context and modality, and users reported that touch modality required more effort than voice modality due to the additional context. Users felt that more effort was needed to use the touch modality due to the additional context, and accordingly, the voice modality would have felt relatively easy. As a result, the modality switching point shifted toward using more voice modality in a longer sentence for the same number of touches.

Also, the change in modality switching point was more prominent in the reading context compared to the watching context. This implies that a combination of physical and mental resource demands may have a greater impact on modality selection than multiple mental resource demands. It also supports what Garde et al. (2002) found in previous studies, that physical demands have a greater impact on the task than mental demands. The cause of this result can be attributed to the change in physical and mental interaction efforts. Although both modalities required more effort in the reading context, the efforts of touch increased more significantly than that of voice, as evidenced by the physical and mental effort graphs. People felt that it took more effort to deal with the overlap of physical resources than the superposition of mental resource demands. This difficulty with the overlap of physical resources probably caused people to use the voice modality more in reading than in watching.

However, the satisfaction results showed a different pattern compared to the results of the interaction efforts. Even though the interaction efforts of voice modality increased, the satisfaction of voice interaction remained unchanged across contexts in this study, indicating a well-known relationship between perceived ease of use and satisfaction with differing results (Calisir & Calisir, 2004). In contrast, for touch modality, satisfaction decreased in an inverse proportion to the increased interaction effort in the reading context. These findings suggest that satisfaction with interactions in multimodal systems is not solely determined by the interaction effort of each modality but rather by their relative superiority.

6. Conclusions

This study developed a multimodal system based on voice and touch modalities and analyzed the effects of their characteristics and context on user modality selection in various usage contexts. The results indicated that the number of S/T determined the modality selection between voice and touch, and the context influenced modality selection by encouraging the use of the voice modality. Furthermore, the developed modality selection prediction model was highly accurate in predicting the dominant modality. These findings contribute to the understanding of the factors that influence modality selection and provide useful insights for the design of more efficient multimodal systems.

This study is significant in several ways. First, it proposes the S/T as a significant variable for comparing voice and touch modalities, enabling quantitative comparison of voice. Moreover, by predicting the modality switching point and the conditions in which each modality is predominantly used, it provides useful insights for multimodal system designers regarding users' modality selection and development directions for the functions they develop. In particular, the study highlights the need to redesign existing interfaces when users' modality selection changes according to the context of use.

The usage trade-offs of voice and touch presented in this study are meaningful, but there may be some discrepancies

between the tasks designed in this study and real-world systems. Real-world systems are composed of a mix of hierarchical and non-hierarchical tasks, and speech is more natural than this study. Therefore, it is necessary to validate this study with real-world systems in future research.

Although many factors were considered in comparing voice and touch modalities, such as the number of S/T and context, the various gestures of touch were not considered in this study. Gestures such as swipe, flick, and pinch in and out may also be used in modality selection, which emphasizes the need for further research on this topic. Additionally, this study used speech stimuli in Korean, and cultural differences may affect the results. Therefore, future research should investigate modality selection in different languages, such as English, Japanese, French, and Spanish, to provide a more comprehensive understanding of the factors that influence modality selection in multimodal systems.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Min Chul Cha  <http://orcid.org/0000-0001-9301-4281>

Yong Gu Ji  <http://orcid.org/0000-0002-0697-2164>

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>
- Baxter, M., Bleakley, A., Edwards, J., Clark, L., Cowan, B. R., & Williamson, J. R. (2021). "You, Move There!": Investigating the impact of feedback on voice control in virtual environments [Paper presentation]. CUI 2021–3rd Conference on Conversational User Interfaces, Bilbao (online), Spain. <https://doi.org/10.1145/3469595.3469609>
- 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 Annual Meeting*, 58(1), 2156–2160. <https://doi.org/10.1177/1541931214581453>
- Budiu, R. (2013). *Interaction cost*. <https://www.nngroup.com/articles/interaction-cost-definition/>
- Calisir, F., & Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Computers in Human Behavior*, 20(4), 505–515. <https://doi.org/10.1016/j.chb.2003.10.004>
- Card, S. K., Mackinlay, J. D., & Robertson, G. G. (1990, April). *The design space of input devices* [Paper presentation]. Conference on Human Factors in Computing Systems–Proceedings, Seattle, Washington, USA. <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>
- Chen, X., & Tremaine, M. (2006). *Patterns of multimodal input usage in non-visual information navigation* [Paper presentation]. Proceedings of the Annual Hawaii International Conference on System Sciences, Kauai, HI, USA. <https://doi.org/10.1109/HICSS.2006.377>
- Cherubini, M., Anguera, X., Oliver, N., & de Oliveira, R. (2009). *Text versus speech* [Paper presentation]. Proceedings of the 11th International

- Conference on Human-Computer Interaction with Mobile Devices and Services, Bonn, Germany. <https://doi.org/10.1145/1613858.1613860>
- Detjen, H., Geisler, S., & Schneegass, S. (2020, October). *Maneuver-based control interventions during automated driving: Comparing touch, voice, and mid-air gestures as input modalities* [Paper presentation]. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada. <https://doi.org/10.1109/SMC42975.2020.9283431>
- Ferrand, L. (2000). Reading aloud polysyllabic words and nonwords: The syllabic length effect reexamined. *Psychonomic Bulletin & Review*, 7(1), 142–148. <https://doi.org/10.3758/BF03210733>
- Findlater, L., & McGrenere, J. (2008). *Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces* [Paper presentation]. Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems—CHI, '08, Florence, Italy. <https://doi.org/10.1145/1357054.1357249>
- Foley, M., Casiez, G., & Vogel, D. (2020). *Comparing smartphone speech recognition and touchscreen typing for composition and transcription* [Paper presentation]. Conference on Human Factors in Computing Systems—Proceedings, Honolulu, HI, USA. <https://doi.org/10.1145/3313831.3376861>
- 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>
- Garde, A. H., Laursen, B., Jørgensen, A. H., & Jensen, B. R. (2002). Effects of mental and physical demands on heart rate variability during computer work. *European Journal of Applied Physiology*, 87(4–5), 456–461. <https://doi.org/10.1007/s00421-002-0656-7>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52(C), 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hauptmann, A. G., & Rudnicky, A. (1990). *A comparison of speech and typed input*. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*. <https://doi.org/10.3115/116580.116652>
- Hoffmann, F., Tyroller, M.-I., Wende, F., & Henze, N. (2019). *User-defined interaction for smart homes* [Paper presentation]. Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia, Pisa, Italy. <https://doi.org/10.1145/3365610.3365624>
- Hua, Z., & Ng, W. L. (2010). *Speech recognition interface design for in-vehicle system* [Paper presentation]. Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications—AutomotiveUI '10, AutomotiveUI, Pittsburgh, Pennsylvania, USA. <https://doi.org/10.1145/1969773.1969780>
- Huang, J., Qi, M., Mao, L., An, M., Ji, T., & Han, R. (2021). User-defined gestures for mid-air interaction: A comparison of upper limb muscle activity, wrist kinematics, and subjective preference. *International Journal of Human-Computer Interaction*, 37(16), 1516–1537. <https://doi.org/10.1080/10447318.2021.1898825>
- 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>
- ISO. (1998). *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs): Part 11: Guidance on usability*. <https://www.iso.org/standard/16883.html>
- Jameson, A., & Klöckner, K. (2005). User multitasking with mobile multimodal systems. In W. Minker, D. Bühler, & L. Dybkjær (Eds.), *Spoken multimodal human-computer dialogue in mobile environments*. Text, Speech and Language Technology (Vol. 28). Springer, Dordrecht. https://doi.org/10.1007/1-4020-3075-4_19
- Jeon, M., Gable, T. M., Davison, B. K., Nees, M. A., Wilson, J., & Walker, B. N. (2015). Menu navigation with in-vehicle technologies: Auditory menu cues improve dual task performance, preference, and workload. *International Journal of Human-Computer Interaction*, 31(1), 1–16. <https://doi.org/10.1080/10447318.2014.925774>
- Kim, J., Jeong, M., & Lee, S. C. (2019). *Why did this voice agent not understand me?* [Paper presentation]. Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings, Utrecht, Netherlands. <https://doi.org/10.1145/3349263.3351513>
- Kim, M. J., Hwangbo, H., & Ji, Y. G. (2020). Comparing flat and edge-screen smartphones operated on a one-hand-only basis: A video observation in laboratory settings. *International Journal of Human-Computer Interaction*, 36(18), 1756–1767. <https://doi.org/10.1080/10447318.2020.1785153>
- Kim, Y., Reza, M., McGrenere, J., & Yoon, D. (2021). *Designers characterize naturalness in voice user interfaces: Their goals, practices, and challenges* [Paper presentation]. Conference on Human Factors in Computing Systems—Proceedings, Yokohama, Japan. <https://doi.org/10.1145/3411764.3445579>
- Kocaballi, A. B., Laranjo, L., & Coiera, E. (2019). Understanding and measuring user experience in conversational interfaces. *Interacting with Computers*, 31(2), 192–207. <https://doi.org/10.1093/iwc/iwz015>
- Laureiti, C., Cordella, F., Di Luzio, F. S., Saccucci, S., Davalli, A., Sacchetti, R., & Zollo, L. (2017). *Comparative performance analysis of M-IqEMG and voice user interfaces for assistive robots* [Paper presentation]. IEEE International Conference on Rehabilitation Robotics, London, UK. <https://doi.org/10.1109/ICORR.2017.8009380>
- Lee, K. M., & Lai, J. (2005). Speech versus touch: A comparative study of the use of speech and DTMF keypad for navigation. *International Journal of Human-Computer Interaction*, 19(3), 343–360. https://doi.org/10.1207/s15327590ijhc1903_4
- 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, 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. In *Pervasive 2008 Workshop Proceedings (Sydney, Australia, May 18, 2008)*. IMUx (improved mobile user experience) (pp. 208–217).
- Lemmelä, S., Vetek, A., Mäkelä, K., & Trendafilov, D. (2008). *Designing and evaluating multimodal interaction for mobile contexts*. [Paper presentation]. Proceedings of the 10th International Conference on Multimodal Interfaces, Chania, Crete, Greece. <https://doi.org/10.1145/1452392.1452447>
- Liu, X., & Thomas, G. W. (2017, May). *Gesture interfaces: Minor change in effort, major impact on appeal* [Paper presentation]. Conference on Human Factors in Computing Systems—Proceedings, Denver, Colorado, USA. <https://doi.org/10.1145/3025453.3025513>
- Malaia, E. A., & Wilbur, R. B. (2020). Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1), 1–16. <https://doi.org/10.1002/wcs.1518>
- McGee, M. (2004). *Master usability scaling: Magnitude estimation and master scaling applied to usability measurement*. [Paper presentation]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria. <https://doi.org/10.1145/985692.985735>
- Mustonen, T., Olkkonen, M., & Häkkinen, J. (2004, June). *Examining mobile phone text legibility while walking* [Paper presentation]. Conference on Human Factors in Computing Systems—Proceedings, Vienna, Austria. <https://doi.org/10.1145/985921.986034>
- Park, S., Kyung, G., Choi, D., Yi, J., Lee, S., Choi, B., & Lee, S. (2019). Effects of display curvature and task duration on proofreading performance, visual discomfort, visual fatigue, mental workload, and user satisfaction. *Applied Ergonomics*, 78, 26–36. <https://doi.org/10.1016/j.apergo.2019.01.014>
- Perakakis, M., & Potamianos, A. (2008). 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>
- Qiu, L., & Benbasat, I. (2005). Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human-Computer Interaction*, 19(1), 75–94. https://doi.org/10.1207/s15327590ijhc1901_6
- Reicherts, L., Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's good to talk: A comparison of using voice versus screen-based interactions for agent-assisted tasks. *ACM Transactions*

- on *Computer-Human Interaction*, 29(3), 1–41. <https://doi.org/10.1145/3484221>
- Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. A. (2018). Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4), 1–23. <https://doi.org/10.1145/3161187>
- Sauro, J., & Dumas, J. S. (2009). *Comparison of three one-question, post-task usability questionnaires* [Paper presentation]. Conference on Human Factors in Computing Systems–Proceedings, Boston, MA, USA. <https://doi.org/10.1145/1518701.1518946>
- Schaffer, S., Jöckel, B., Wechsung, I., Schleicher, R., & Möller, S. (2011, August). Modality selection and perceived mental effort in a mobile application. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Interspeech 2011, Florence, Italy. <https://doi.org/10.21437/Interspeech.2011-599>
- Schartmüller, C., & Riener, A. (2022). *Multimodal error correction for speech-to-text in a mobile office automated vehicle: Results from a remote study* [Paper presentation]. International Conference on Intelligent User Interfaces, Proceedings IUI, Helsinki, Finland. <https://doi.org/10.1145/3490099.3511131>
- Sendlneier, W. F. (1995). Feature, phoneme, syllable or word: How is speech mentally represented? *Phonetica*, 52(3), 131–143. <https://doi.org/10.1159/000262128>
- 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>
- Suhm, B., Myers, B., & Waibel, A. (1999). *Model-based and empirical evaluation of multimodal interactive error correction* [Paper presentation]. Conference on Human Factors in Computing Systems–Proceedings, Pittsburgh, Pennsylvania, USA. <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. (2021). Typing on a smartwatch while mobile: A comparison of input methods. *Human Factors*, 63(6), 974–986. <https://doi.org/10.1177/0018720819891291>
- Wechsung, I., Naumann, A., & Möller, S. (2010, October 1–2). The influence of the usage mode on subjectively perceived quality. In *Spoken Dialogue Systems for Ambient Environments: Second International Workshop on Spoken Dialogue Systems Technology, IWSDS 2010, Gotemba, Shizuoka, Japan. Proceedings* (pp. 188–193). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16202-2_20
- Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3), 449–455. <https://doi.org/10.1518/001872008X288394>
- Wu, J., Chang, C. C., Boyle, L. N., & Jenness, J. (2015, January). Impact of in-vehicle voice control systems on driver distraction: Insights from contextual interviews. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 1583–1587. <https://doi.org/10.1177/1541931215591342>
- Yoon, S. H., Lee, S. C., & Ji, Y. G. (2021). Modeling takeover time based on non-driving-related task attributes in highly automated driving. *Applied Ergonomics*, 92, 103343. <https://doi.org/10.1016/j.apergo.2020.103343>
- Zhao, M., Cui, W., Ramakrishnan, I., Zhai, S., & Bi, X. (2021). *Voice and touch based error-tolerant multimodal text editing and correction for smartphones* [Paper presentation]. The 34th Annual ACM Symposium on User Interface Software and Technology, Virtual Event, USA. <https://doi.org/10.1145/3472749.3474742>
- Zijlstra, F. R. (1993, January). *Efficiency in work behaviour: A design approach for modern tools* (pp. 1–186). Delft University Press. <http://www.csa.com/partners/viewrecord.php?requester=gs&collection=TRD&recid=N9516953AH>

About the authors

Min Chul Cha is a Postdoctoral Researcher in Industrial Engineering at Yonsei University, Seoul, Korea. He received his PhD from Yonsei University in 2023. His research interests include multimodal interface, voice user interaction, automotive user interface, and smart devices.

Yong Gu Ji is a Professor in the Department of Industrial Engineering at Yonsei University, where he directs the Interaction Design Laboratory. He received his PhD in Human Factors/HCI from Purdue University. His research interests include usability/UX in smart devices and self-driving vehicles.

Appendix A

Subjective Questionnaires

1. How much mental effort was required to perform the tasks in each modality?

1-1. Voice (exclude touching the PTT button)



1-2. Touch

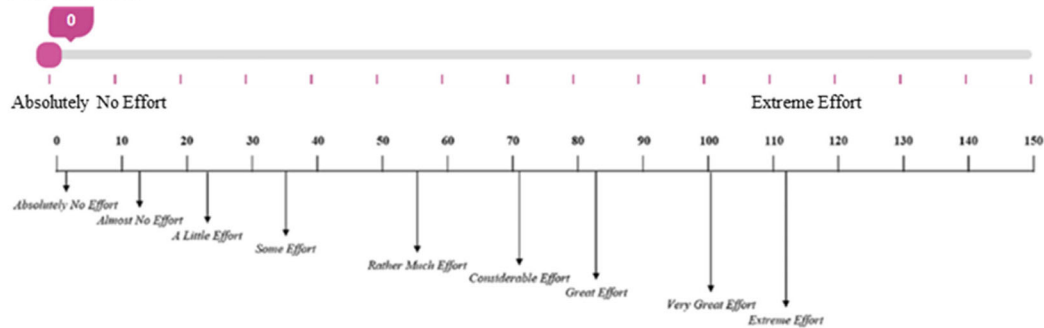


2. How much physical effort was required to perform the tasks in each modality?

2-1. Voice (exclude touching the PTT button)



2-2. Touch



3. How satisfying was it to perform the tasks in each modality?

3-1. Voice (exclude touching the PTT button)



3-2. Touch

