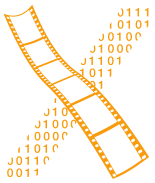

Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

I can say "Navigate me there": A Comparison of Different Interface Designs to Improve the Discoverability of Multimodal Interaction Techniques

Master's Thesis
submitted to the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Yeganeh Sadat Hajimiri

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Enrico Rukzio

Registration date: 25.6.2021
Submission date: 22.12.2021

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

Hajimiri, Yeganeh Sadat | 383676

Name, Vorname/Last Name, First Name | Matrikelnummer (freiwillige Angabe) Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare in lieu of an oath that I have completed the present paper/Bachelor thesis/Master thesis* entitled

I can say “Navigate me there”: A Comparison of Different Interface Designs to Improve the Discoverability of Multimodal Interaction Techniques

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

Aachen, 22.12.2021 | _____

Ort, Datum/City, Date | Unterschrift/Signature

*Nichtzutreffendes bitte streichen

*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
Para. 156 StGB (German Criminal Code): False Statutory Declarations
Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence
(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.
(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Aachen, 22.12.2021 | _____

Ort, Datum/City, Date | Unterschrift/Signature

Contents

List of Figures

List of Tables

Abstract

The research and technology in the field of multimodality are becoming mature enough to let us use multimodal interactions in daily applications and benefit from the advantages of a multimodal system, such as increasing efficiency, context flexibility, and improving user satisfaction. But the problem is that there exist design gaps, like discoverability of the new multimodal interactions on an interface, that are not comprehensively investigated in the literature. For example, in most studies, the users have been told about the possibility of multimodal interactions on an interface and how they can be performed. But the fact is that we are not with the users in a daily application when they start using this new class of interface. Therefore, in this thesis, we propose different interface designs to improve the discoverability of *gaze-speech* and *touch-speech* multimodal interactions on a map application to do location-related tasks. Experts helped us to find the two most promising interface designs (*InteractionMap* and *Game*). We evaluated these two interfaces based on the discoverability, awareness, and learnability of the two target multimodal interactions. In a between-subject user study with 36 users, a comparison of the two interfaces with a baseline, which introduces the new interactions using a video, shows that the two interfaces are not different from the baseline. In fact, *Game* was better than the other two in *touch-speech* awareness, and very similar to *Baseline* in *gaze-speech* awareness. The findings of this thesis can help future works to enrich the multimodal interactions in daily applications for novices.

Acknowledgements

This thesis was in collaboration with Mercedes-Benz AG. It was a privilege to work with Dr. Felix Schüssel from Mercedes-Benz and Oliver Nowak i10. They patiently listened to my ideas and thoughts every week, and guided me through an instructive and enjoyable thesis process.

I would like to thank all my wonderful colleagues at Mercedes-Benz for voluntarily participating in my user studies and giving me valuable feedback.

It is my pleasure to thank Prof. Dr. Jan Borchers and Prof. Dr. Enrico Rukzio for examining this thesis.

Additionally, I would like to thank my parents, who always gave me their blessings.

Finally, I would like to thank my boyfriend, Mostafa, for all his unwavering support.

Thanks you!

Yeganeh Hajimiri

Conventions

Throughout this thesis we use the following conventions.

Text conventions

This thesis is written in American English. The first person is written in the plural form. Unidentified third persons are described in female form.

Chapter 1

Introduction

According to Oviatt [2007] and Sebe [2009], a multimodal system is a system that processes the combination of more than one input modality or communication channel. As Oviatt [2007] mentioned, in the recent multimodal interfaces, the input mostly means human-like modalities like speech, gaze, or gesture. Moreover, Oviatt [2007] named the recognition technology to recognize each input modality, and the architecture for semantic integration of the recognized modalities as parts of the process for finding the intention of a multimodal interaction in multimodal systems. In addition to works about the technology and the architecture of a multimodal system, there are also more HCI-related works in the literature. Some of these HCI works investigate on benefits, such as efficiency or consistency, of using multimodal interactions in different applications [Cohen et al., 2000, Srinivasan et al., 2020]. And as Oviatt [1999] explained in one of the *myth* from her *ten myths* of multimodal interactions, the benefits of these interactions are not only efficiency, but they can also be context flexibility or user satisfaction. In addition, according to one other *myth*, if a multimodal interaction on an interface does not feel natural for the users, they would not necessarily interact multimodally. Therefore, some other works in HCI focus on finding the natural multimodal interaction for a task [Schüssel et al., 2013].

Human computer interaction can benefit from multimodal interactions.

However, the users' previous experiences with a system

Users are habituated to touching the displays even if they benefit from other input modalities.

could affect their choice of modality. For example, when a display interface is given to the users, the first input attempt could be touch [Schüssel et al., 2013, Srinivasan et al., 2020]. The authors assume that the reason behind these findings is that the users are habituated to interact with a display interface using touch. If we do not provide clear signifiers about the existence of multimodal interaction, especially in a familiar interface for users, we cannot expect them to use that interaction. Therefore, the problem is how users would discover a multimodal interaction on an interface, so they would be aware of it and learn how to perform it for the tasks that multimodality may have advantages.

Discoverability of a new multimodal interaction is done by an instruction in the literature.

In addition, in the works that examine a multimodal interaction for a specific context, Srinivasan et al. [2020] gave users a complete explanation over the new multimodal interaction as part of the study instruction, and Cohen et al. [2000] continued the instruction with a practice session to make users experts in the target interaction. Therefore, a research gap is how to design a multimodal interface that could make its supported multimodal interactions discoverable and take the place of personally given instructions in the real world.

Discoverability of multimodal interactions is a design research gap.

To our knowledge, discoverability is a design challenge that is not well investigated for multimodal interactions. In fact, the design of a multimodal interface, in general, is not widely explored, and there is no detailed design guideline that can be followed. A primary reason for this design challenge could be different specific design decisions that, according to Reeves et al. [2004], the variety of multimodal interaction techniques, their context of use, including user, task, and application, require.

We investigated the ideas for discoverability of two multimodal interaction techniques.

The goal of this thesis is to find the approaches that improve the discoverability of two multimodal interaction techniques (*gaze-speech* and *touch-speech*) on a map application to do location-related tasks. We got the help of experts to find the two best approaches among our seven initial ideas. One approach is about providing a mapping of the interactions, including real-time feedback on what is detected and what is possible (*InteractionMap*). The other approach introduces the new multimodal interactions in

a game context at the beginning of the usage (*Game*). We compared the discoverability for these two candidate approaches with a *Baseline*, which introduces the new interaction techniques at the beginning like an instruction.

1.1 Outline

In the following, we review the related works in Chapter 2. We talk about the human factors for learning new techniques, and works in the literature on discoverability of different interaction techniques.

Chapter 3 "Designing Discoverable Multimodal Interfaces" is about the designs that we proposed to improve the discoverability of *gaze-speech* and *touch-speech* multimodal interactions. We first describe the multimodal system of our focus. After that, we explain the ideas behind the design of each *proposed interface* and the detail of their appearance and behavior. In the end, we report the expert study in Section 3.3 that we conducted to evaluate these interfaces and pick the two *candidate interfaces* for further evaluation.

In Chapter 4 "A Comparison of Different Interface Designs for Better Discoverability", we describe the evaluation study we conducted to compare the discoverability of the two multimodal interactions for two *candidate interfaces* (*InteractionMap* and *Game*), with the *Baseline*. Furthermore, we discuss factors in these interfaces that contribute to or hinder discoverability in Section 4.3.9.

We further discuss general factors that could influence the discoverability of multimodal interactions in Chapter 5 "Discussion".

In Chapter 6 "Summary and future work", we summarize the results of this thesis and suggest future research regarding the discoverability of multimodal interactions.

Chapter 2

Related work

2.1 Ideas for Discoverability

Achieving a new skill includes three phases: *cognitive*, *associative*, and *autonomous*.

Based on the psychology literature, Cockburn et al. [2014] talked about the different phases of skill acquisition: *cognitive*, *associative*, and *autonomous*. The *cognitive* phase is about awareness of possibilities. Different methods, like instruction, clear models, or some feedback, can be used for this learning phase. In the next phase, the *associative* phase, learning continues with finding how things are done. In addition, learning many tasks might take many years to go to the *autonomous* phase, in which the skill can be performed automatically. Furthermore, they mentioned some approaches for improving a skill for physical activities that can be used for improving skill in computer interfaces, for example, *power low of practice* [Fitts and Posner, 1967], *motivation*, *feedback*, and *guidance*. In this thesis, our goal is discoverability of an interaction technique. According to the works of Chueke et al. [2017], Walter et al. [2013], and Goguey et al. [2018], discoverability of an interaction mainly aims for two first phases of skill acquisition.

Discoverability is about first two phase of learning: *cognitive (what)* and *associative (how)*

Moreover, the main part of the work of Cockburn et al. [2014] was grouping the works on skill acquisition in computer interfaces into four categories: *intramodal improvement*, *intermodal improvement*, *vocabulary extension*, and *task mapping*.

Some approaches are used for improving the usage of a single interaction technique.

Works under the *intramodal improvement* category concentrated on improving the performance of one interaction technique in the skill acquisition phases. For example, Freeman et al. [2009], Bau and Mackay [2008] used *feedforward* as a *guidance*, and Kurtenbach and Buxton [1993] provided *rehearsal* to novices for discovering gestural interactions and passing through the *associative* learning phase. In addition, suggestions from game applications by Dyck et al. [2003] to have *transient text*, *audio*, and *animation* in a non-distracting manner were for the *cognitive* phase of learning in this category.

Different approaches can be used to invite users to a new alternative interaction.

The *intermodal improvement* category included those works that assist users in switching to a more efficient interaction technique. In some cases, the transition to a new interaction technique was susceptible to *dip performance*. One approach to reach *intermodal improvement* included *forcing* users to the new interaction. For example, Grossman et al. [2007] used dwelling on the old technique to let users observe the hints for the new technique, which was disruptive. Furthermore, the other approaches were to give a nonintrusive *recommendation* on the new interaction [Scarr et al., 2011] or to give *motivation* by reporting on users' performance when they switch to the new interaction [Malacria et al., 2013].

Scarr et al. [2011] worked on the advanced interaction techniques on an interface that are less usual than the routine interaction techniques. They introduced a framework of the influencing factors in reaching the expert performance in using an interface. For example, some of these factors that are helpful for this thesis and should be kept in mind are: the new interaction should be visible and ready-to-hand, the vocabulary learning should be incidental, and users tend to stick to the familiar interactions. They introduce *Blur* which is a system that follows the framework to initiate switching from a WIMP interaction to the advanced command-line interaction. This system includes two parts *calm notification* and *hot commands*. *Calm notification* (Figure 2.1) is the introduced approach to take care of the three above-mentioned factors of the framework and tries to promote learning and awareness for the *hot commands*. This system showed a high performance in switching to the advanced interaction.

Figure 2.1: The default state of the *Blur* system (top). This system aims for initiating a switch to the command-line interaction. A *calm notification* appears when an operation is done by WIMP interaction and it suggests the command name for the alternative command-line interaction (bottom). Images are taken from [Scarr et al., 2011].

Recommendations are a way to extend users' knowledge over new possibilities.

Works under the *vocabulary extension* category focused on increasing the users' knowledge of available functionalities on the interface. However, some approaches in this category were similar to previous categories. The approaches were to provide recommendations and suggestions. Cockburn et al. [2014] concluded that the *recommendations* should be generated and presented according to the context of use, and they should not disrupt users from the current task.

Gamification is helpful for developing strategy.

The last category, *task mapping*, included works that try to help users to develop strategies for doing a task in the user interface. Using *gamification* was one way to provide strategical training [Li et al., 2012].

In addition, there are works on discoverability of speech commands using different approaches like introducing the command possibilities as a walkthrough at the beginning [Feng et al., 2005], or recommending commands based on the current context of use [Furqan et al., 2017, Srinivasan et al., 2019].

Srinivasan et al. [2019] suggested three interface designs for discovering natural language commands in a touch-speech multimodal photo editing application using a contextual-

example recommendation approach. Their suggested ideas for these three interfaces differed in where, when, and what contextual-example to suggest. Their first idea (*exhaustive*) was showing examples for all the possible operations on a fixed location when the tap-to-talk button is pressed. However, the examples for irrelevant operations to the current state of the application were grayed out. The second idea (*adaptive*) was to show the context-relevant examples as an overlay on the finger's location when the user long presses on the application buttons, the canvas, or the objects. The last idea (*embedded*) was similar to the previous idea, except the suggestions were shown alongside the application GUI instead of overlying on the screen. Furthermore, when an error occurred, a feedback message on a fixed location was shown. This feedback included an example command in addition to the error message. Their results showed that the *adaptive* idea was more encouraging for users to use speech input, and also, the feedback message could be effective in improving the discoverability.

When suggestions overlay where the interaction happens, they are more encouraging.

2.1.1 Design Principles

Feedback, *feedforward*, and *affordance* are helpful design principles for discoverability.

Norman [1988], Djajadiningrat et al. [2002], Hartson [2003] and some others defined *feedback*, *feedforward*, and *affordance* in their works. Vermeulen et al. [2013] clarified the definition of these design principles in the literature. *Feedforward* tells the user the purpose or the result of an action, aiming to bridge Norma's Gulf of execution. On the other hand, *affordance* is the prerequisite of the *feedforward* and invites users to take the proper action. They introduced four classes for *feedforward*: *false*, *hidden*, *nested*, and *sequential*. In *nested feedforward*, the purpose of action can be conveyed through a different level. For example, at a high level, the whole system can tell the purpose of a button. Moreover, *sequential feedforward* is for actions that logically follow each others; therefore, *feedback* of the first action turns to *feedforward* for the second action. A *feedforward* is *false* or *hidden* when it is incorrect or misleading in expressing a functionality in the system.

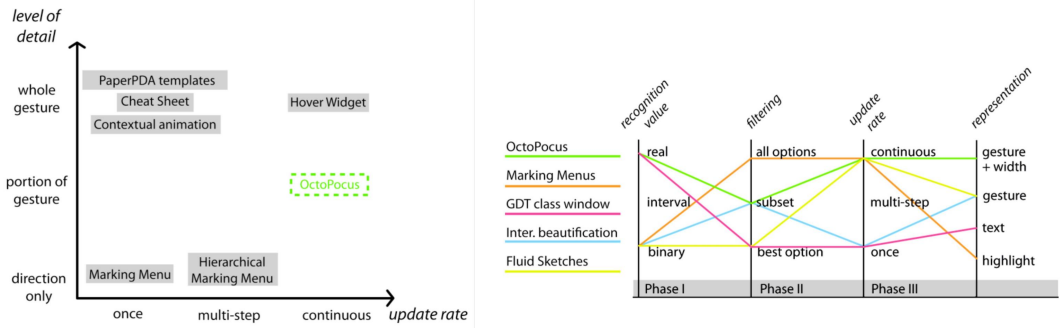

Figure 2.2: Design space for feedforward (left) and feedback (right) in learning new gestural interactions. The pictures show the identified dimensions for feedforward and feedback and the possible values in each dimension. Images are taken from [Bau and Mackay, 2008].

Bau and Mackay [2008] identified a design space for feedback and feedforward in the research area of learning new single-touch gestures. They discussed *level of detail* and *update rate* as the two dimensions for designing feedforward in this area (Figure 2.2(left)). In addition, *recognition value*, *filtering*, *update rate*, and *representation* were described as the different dimensions for feedback in this design space (Figure 2.2(right)). Freeman et al. [2009] suggested the *degree of co-location* as an additional dimension for this design space. They discussed that the the learning of a new gesture can happened either in-situ or in a separate mode.

Gaze interaction can improve the application of the design principles.

Aslan et al. [2018] proposed gaze-triggered *affordance* for graphical user interface elements. Mouseover tooltips was the inspiration of their approach. By looking at a graphical user interface element, like a switch or a knob, for a short time, an animation would reveal the possible movement for that element. They compared this *gazeover* idea with the traditional mouseover and found that the *gazeover* is more enjoyable than mouseover.

Mapping and *conceptual model* are good approaches for creating mental model in complex systems.

Feedforward, *feedback*, and *affordance* are helpful design principles. However, as Norman [1988] mentioned in his book, in complex systems, *mapping* and *conceptual model* are two ways to convey the purpose of an interface. *Mapping* shows the relationship between the control and its effects. A *conceptual model* helps users to understand how things work

in a preferably simplified way. They both try to create a helpful mental modal for users to allow them to utilize the system.

2.2 Discoverability Evaluation

Apart from the ideas and design principles to design for discoverability, part of the literature review for this thesis included understanding how to conduct a study to evaluate discoverability.

Walter et al. [2013] designed for discoverability of a mid-air gesture on public display while playing a game. Using the location and timing of a hint, they provided three ideas. The first idea was to interrupt the game by showing the hints (*temporal division*). The second idea was to dedicate a permanent area of the screen to the hints (*spatial division*), and the third idea was to integrate the visual hints into the game itself (*integration*). They differed in how they provided hints for the first and second ideas: *text*, *text plus icon*, and *text plus video*. In the study design, they gave participants time to do the game without providing any information on the game and the gesture. The study would end if participants could perform the gesture or after 2 minutes. Moreover, they did an interview to find answers to their research questions: whether or not participants noticed the hints, performed the gesture, and understood the gesture before trying. Their result showed that the percentage of the users who performed the gesture was highest for *temporal division*, there is no difference between different types of hint, and if users understand the gesture, they probably perform it.

Goguey et al. [2018] introduced two force-sensitive text-selection techniques and designed visual *feedback* for discoverability of these techniques. They aimed to improve expert performance ceiling in text-selection tasks. Furthermore, they designed visual cues to ease the discovery of these techniques for the novices. In the study to evaluate discoverability, they did not explain the technique to the participants. However, they explained the possibility of

force on the touch screen and the registration interactions like triggering the selection by long press. Then, participants performed two blocks of nine text-selection tasks after going through an exploration phase with no time limit. At the end, they were asked to explain the technique. In this study, the important variables were the exploration time, whether the participants could figure out the technique, and whether they could explain it.

Hofmeester and Wolfe [2012] used an iterative design process to design a discoverable design for *swipe to select* gesture. In addition, they used the RITE methodology to apply the necessary changes quickly between sessions of each iteration. In the studies they conducted, they first gave users some time to explore the interface, and then without giving any hint on the interaction, they gave them some related tasks. The questions they tried to answer were about: how users use the interaction, how it affects their experience, how confident they feel about it, how fast they discover and learn it, how they describe it, and can they use the knowledge they get across the system. However, we did not follow their design process in this thesis; their evaluation questions inspired this thesis.

2.3 Multimodal Interface Design

Design a multimodal interface differs from a standard interface.

According to Turk [2014], designing a multimodal interface is challenging since many of of the design guidelines for standard user interfaces do not apply for them. Moreover, many of design decisions in a multimodal system would change depending on the application, tasks, context of use, and involved multimodal interaction techniques.

There are some guidelines for multimodal interface design.

Reeves et al. [2004] introduced six categories of guidelines for the multimodal interface design: *requirement specifications*, *designing multimodal input and output*, *adaptivity*, *consistency*, *feedback*, and *error prevention/handling*. Some of the points in these categories are beneficial for our work. The guideline says a multimodal interface should not force users to a specific modality or, in general, multimodal interaction. Also, users should be aware of the interaction

Figure 2.3: Affordance and feedback for multimodal interactions in a visualization application. By pointing at an axis, the other possibilities, like speech (command box) and pen interaction (ink pad), are highlighted. Image is taken from [Srinivasan et al., 2020].

that they are performing and the possible modalities. However, this should be done using a descriptive form instead of lengthy instruction that might distract the users from the task. In addition, confirmations should not be on each modality rather on the system's whole interpretation.

Srinivasan et al. [2020] introduced consistent multimodal interaction techniques using pen, touch, and speech for data visualization applications. Discoverability of the interactions was not their goal, and users were informed at the beginning of the study about the multimodal possibilities and practiced them in an unlimited time frame. However, their results show that there were still participants who preferred touch interaction. In addition, they got the help of *affordance* and *feedback* to help users correctly perform the interactions and recover from errors. Based on the user's first action, they highlighted the possibilities that exist, like the speech command area for speech interaction and ink pad area for pen interaction, to perform a valid interaction (Figure 2.3). They also used *feedback* to communicate the result of a completed interaction, whether *successful*, *invalid*, or *void*. They discuss that these design elements were not

noticeable in most cases.

2.4 Multimodal Interaction Techniques

In order to better design for the discoverability of multimodal interactions, it is important to know the characteristics of these interactions.

Oviatt [1999] identified and accurately modified ten *myths* for creating a multimodal system and some of them can help us with our goal for this thesis. *Myth*: "If you build a multimodal system, users will interact multimodally." However, this behavior towards performing multimodal interactions depends on the action type. An action type predictably determines whether users choose multimodal or unimodal interactions to perform it. *Myth*: "Multimodal input involves simultaneous signals." These simultaneous signals rarely overlap and mostly occur sequentially. For example, in a multimodal interaction using touch and speech as the input signals, users would mostly perform them sequentially.

Probability of switching to multimodal interaction depends on the task type.

Oviatt et al. [1997] worked on finding natural interaction patterns in multimodal interactions using pen and speech on a map application. Their first finding was that the user's probability of doing a task multimodally depends on the command type needed for the task. Their results show that tasks that involved *spatial location commands*, which included specifying spatial location about a point, were performed multimodally more often than tasks with *selection commands*. In the latter command type, specifying the object was easy in the spoken form, and therefore performing a gestural interaction to specify the object felt unnecessary. As Oviatt [2007] discussed, users would shift to multimodal interaction if the task gets complex and increases the mental load.

Chapter 3

Designing Discoverable Multimodal Interfaces

In this chapter, we describe our journey towards the two multimodal interfaces (*candidate interfaces*) we investigate in Chapter 4 "A Comparison of Different Interface Designs for Better Discoverability". For this, we start with explaining the look and feel of the multimodal base system (*base system*) of our focus. Then, we describe our seven proposed interface designs for improving the discoverability of multimodal interaction (*proposed interfaces*) and their specific appearances and behaviors. In the end, we conduct an expert study and discuss the findings to pick the two *candidate interfaces* from the *proposed interfaces* for the final evaluation.

3.1 Look and Feel of the Multimodal Base System

Description of the *bases system*.

Different specifications, like tasks, interaction techniques, and underlying logic, define a multimodal system. In the following, we describe the specifications of the *base system* for which we try to improve the discoverability of its multimodal interactions in this thesis.

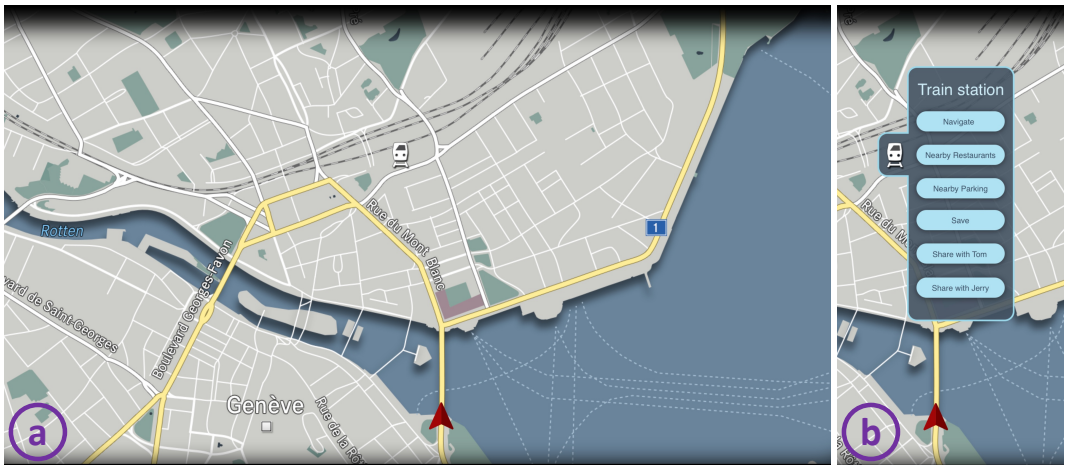

Figure 3.1: Map application as our *base system*(a). Menu on the train station which shows the possible tasks for a location (b).

3.1.1 Multimodal Interaction Techniques, Applications, and Tasks

The *base system* is a map application supporting location-related tasks.

Multimodal interaction is a natural behavior of humans, and they have a lot of multimodal experiences in their human-human interactions. However, when it comes to multimodal interaction on a computer interface, the intuitiveness of the multimodal interaction for the target context should be examined [Reeves et al., 2004]. Multimodally interacting with a map application has shown efficiency and naturalness in literature [Cohen et al., 2000, Oviatt et al., 1997]. As such, we have picked a map as our application, and the possible tasks on this map application are familiar tasks on a location, similar to Google Maps. Figure 3.1 shows the map application as it is used throughout this thesis and the six possible tasks for a location, which is a train station in our application.

The *base system* supports *gaze-speech* and *touch-speech* multimodal interaction techniques.

Similar to Koons et al. [1998], the multimodal interaction techniques that our map application supports for doing the tasks are *gaze and speech* (*gaze-speech*) and *touch and speech* (*touch-speech*). In these two multimodal interaction techniques, users can select a location (train station) using gaze or touch and give a specific command using speech. In addition to these two interaction techniques, our map application supports *touch-only* interaction, which is the primary

interaction technique in all touch interfaces, and *speech-only*, which is a fundamental interaction when an interface supports speech.

Our goal is to improve the discoverability of the two new multimodal interactions on an interface that still supports users' habitual interactions. Therefore, all the six location-related tasks on this map application are possible using all four supported interactions. Moreover, these tasks are all integrated into the map application in a way that users can perform them within 2-steps in a *touch-only* interaction as the primary form of interaction on a touch screen. For example, to navigate to the train station, users can touch the train station to open the menu (shown on Figure 3.1(b)) and then touch the navigate button.

Using a *touch-only* interaction, all tasks on the map application are possible with 2-steps.

3.1.2 Realization of Multimodal Interaction

Besides the specifications mentioned above, our application should follow an underlying logic in order to realize as a comprehensive multimodal system to serve as the *base system* for the *proposed interfaces* we describe in Section 3.2 "Design for Discoverability". The ways that the underlying logic is defied would affect the user experience with the multimodal interface. Therefore, to minimize the effects of extraneous factors on the discoverability of our *proposed interfaces*, we aim to control our *base system* by keeping this system as similar as possible to a touch-only interface. In the following, we explain the logic behind the *base system*.

An underlying logic is needed to realize a multimodal interaction.

Fusion. A fusion architecture is the background of every multimodal system. It defines how to process input modalities, depending on their type, and integrate them to a complete and meaningful output as the user's intention. Since our modalities (gaze and speech in *gaze-speech* interaction, and touch and speech in *touch-speech* interaction) are not temporally coupled, we use the *late semantic fusion* approach [Oviatt, 2007]. In this approach, different input modalities are recognized separately and integrated with an understanding component. The integration pro-

In the *base system*, a late semantic fusion architecture is used.

cess can be complex. For example, a speech-based multimodal system uses context and dialog management information and includes alternating lexical candidates to reach the best interpretation. However, since our focus is not on the performance of the fusion architecture, we simplified the integration process for our *base system*.

The fusion of the *base system* starts with recognizing each of the three modalities (gaze, touch, and speech) separately. To recognize speech input, we used an internal tool at Mercedes-Benz named *VoiceConnect*. This tool uses the model created in *Cerence Studio* to recognize the intent of the speech. To recognize the gaze modality, we used another internal tool named *GazeConnect*. This tool gets the gaze data from an eye tracker attached to a display and converts the data to the valid coordinate points for the display. Using *ProtoPie*, a prototyping tool that we used to create the prototypes of the interfaces, we decide on which point of interest (POI) the user is looking at. On our map application, the locations, such as the train station, are the POIs. Also, for the touch modality, *ProtoPie* detects when a POI is touched. Then, another internal tool, *FusionConnect*, serves as an understanding component to integrate the intents of all the recognized modalities. This tool uses a 2 seconds temporal gap [Oviatt et al., 1997] to combine gaze with speech or touch with speech. This simplified integration process is independent of the order that the modalities are received. In case of successful integration, the output is sent back to the *ProtoPie* as the user's overall intention, and *ProtoPie* acts upon meaningful intentions. Furthermore, if users open the menu of a POI by touching it and give the speech command but not within 2 seconds, the temporal fusion would not occur. However, we consider this interaction a *touch-speech* multimodal interaction.

The understanding component is simplified to use a temporal approach to integrate different modalities.

Feedback on inputs. As Reeves et al. [2004] discussed in their multimodal interface design guideline, all the modalities should be coherently connected to represent the interaction state. Therefore, we removed all the feedback on inputs and design for feedback on individual modalities as part of the whole design for multimodal discoverability. For example, we removed the speech wave for the speech

The input feedback on the *base system* is same as a *touch-only* interface.

modality from the *base system*, and it does not include any default feedback on gaze.

Another important feedback that needs to be controlled but cannot be removed is on touch modality. In current map applications on touch screens, one feedback on touching a location is immediately opening a menu. In this thesis, we refer to this feedback by *menu open on touch immediately* (MOT). Since *MOT* is necessary feedback for *touch-only* interaction, and we do not want to change the look and feel of a familiar touch-only system, we keep *MOT* unchanged. However, the *MOT* can be misleading for the users that are discovering the *touch-speech* interaction. For example, if the intention is *touch-speech*, the *MOT* as an affordance for *touch-only* interaction would be unexpected and can harm the discoverability of the *touch-speech*. This challenge results from an immature fusion architecture to detect the user's exact intention and act based on that. Despite this challenge with *MOT*, we decided to keep this feedback in our *base system*.

Modality of system output. As another guideline for multimodal interface design, a multimodal interface should not demand users to comprehend different modalities simultaneously to understand the system's output [Kalyuga et al., 1999]. However, the system's output can be provided in different modalities depending on the user's preferences, the context of use, and system functionality. Since these interrelations are out of the scope of this work, we decided to stick with visual-only output, as found in a *touch-only* interface.

The *base system* only provides visual outputs.

3.2 Design for Discoverability

In this section, we first explain what we want to achieve by designing a discoverable multimodal interface. After that, we describe the *proposed interfaces* by giving details on why we came up with these designs and how they exactly look and behave.

3.2.1 Discoverability

Awareness (what) and learnability (how) are the two key attributes of the discoverability of an interaction.

In our work, discoverability refers to the discoverability of interaction possibilities on an interface. To answer what contributes to a discoverable interaction, we are inspired by the discoverability works in gesture interactions. Chueke et al. [2017], Walter et al. [2013], Goguey et al. [2018] refer to a discoverable interaction not only when it is performed successfully by the users but also when they comprehend it. Since our goal is to improve the discoverability of *gaze-speech* and *touch-speech* interactions in our multimodal *base system*, we define discoverability by two attributes: learnability and awareness. Learnability means the users can learn the interaction in order to perform it. Awareness means the users are mindfully aware of the interaction, and they do not perform it by accident.

3.2.2 Proposed Interfaces

Proposed interfaces are designed to improve the discoverability of the multimodal interactions in the *base system*.

Inspired by literature and doing brain storming, we came up with seven *proposed interfaces* for discoverability of the two multimodal interactions (*gaze-speech* and *touch-speech*) on the *base system*.

Interface 1 and *interface 2* are minimal interfaces with only feedback on gaze.

Interface 1. The goal of this interface is to minimize the design for multimodal discoverability to only feedback on gaze modality because this modality is the only new interaction. The gaze feedback is also minimized to only increasing the size of a POI when it is looked at [Miniotas et al., 2004].

Interface 2. Same as the previous interface, this interface also minimizes the design for multimodal discoverability to only feedback on gaze modality, but with constant gaze feedback.

Interface 3. Inspired by Vermeulen et al. [2013] on feedforward and based on Norman [1988] suggestions on using mapping and conceptual model for discoverability in complex interfaces, we did a brainstorming session to come up with a simple graphical design (*InteractionMap*) that models multimodal interactions. As shown in Figure 3.4(a), the graphical design is placed in the corner of the interface.

Interface 3 uses the mapping to show the possible interactions.

Interface 4. Since multimodal interaction is a natural behavior in human to human interactions [Oviatt, 2007], in this interface, we aimed at evaluating discoverability when the users are only aware of the individual supported modalities on an interface, in our case, gaze, touch, and speech. Therefore, this interface only shows the available modalities without any feedback on them.

Interface 4 is a minimal interface that only provides awareness of possible modalities.

Interface 5. Based on the work of Vermeulen et al. [2013] on sequential feedforward and affordance, we designed this interface. This interface gives users in place hints over what they should do next based on their first action. For example, if a user looks at a POI, a speech icon appears on it to tell the users they can say something about that POI. This interface favors the sequential approach to multimodal interaction; however, according to Oviatt [2007] people show two different highly consistent behaviors towards multimodal interactions. They are either "simultaneous integrators" or "sequential integrators." Therefore, this interface can be difficult for "simultaneous integrators", who tend to combine all the involved modalities at once.

Interface 5 uses the idea of sequential feedforward and affordance.

Interface 6. In the interface, we utilized the playful aspect of the multimodal interactions. We designed a familiar game (*coloring game*) and instructed users to do the game using the two multimodal interactions. However, our idea does not fall entirely into gamification definition by Deterding et al. [2011], we wanted to use the playfulness of a familiar game to increase users' engagement in learning the interactions [Cockburn et al., 2014]. Also, according to survey by Cockburn et al. [2014], this interface provides the

Interface 6 teaches the multimodal interactions throughout a game.

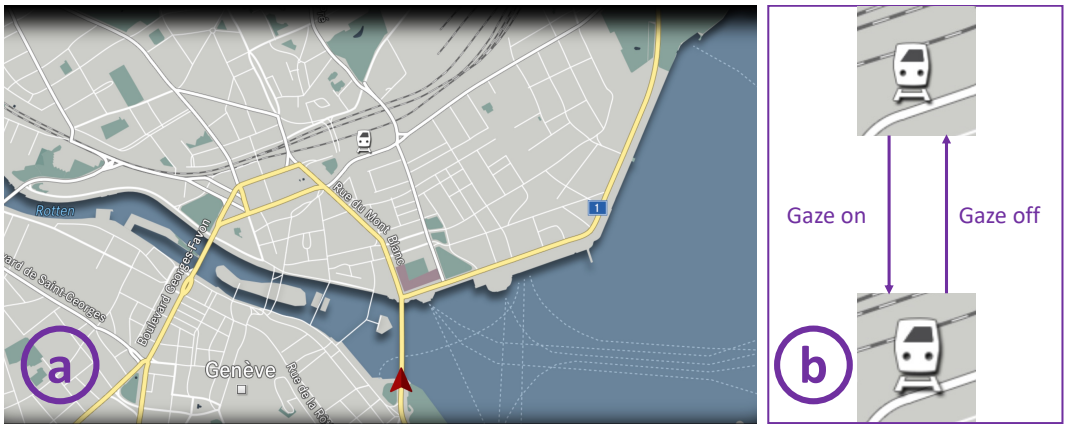

Figure 3.2: *interface 1*, the appearance of this interface is same as *base system*(a). It only has additional *POI size* gaze feedback (b).

users some practice over the interactions and teaches them the interactions as a strategy to do tasks in general and not as a way to do a specific task.

Interface 7 recommends the multimodal interactions in case of detecting an unimodal interaction.

Interface 7. One approach to design a discoverable multimodal interface is to use the idea of "vocabulary extension" by Cockburn et al. [2014] and recommend the two new interactions in the cases that users perform the habitual unimodal interactions (*touch-only* and *speech-only*). Similar to "calm notification" by Scarr et al. [2011], we have designed a window that appears temporally on the top of the screen with information about the two new interactions without interrupting users from their tasks.

3.2.3 Interfaces Appearances and Behaviors

The ideas of *proposed interfaces* are applied to the *base system*, explained in Section 3.1 "Look and Feel of the Multimodal Base System". Furthermore, these ideas include some details to create a coherent interface. In Table 3.1, we list these details for each *proposed interface* including pictures of them.

Proposed Interface	Details of Design
interface 1 (Figure 3.2)	• By looking at a POI, its size increases (*POI size* gaze feedback). Figure 3.2(b) • There is no feedback on the states of the voice recognition.
interface 2 (Figure 3.3)	• A cursor constantly moves with the gaze position (*Cursor* gaze feedback). Figure 3.3 • There is no feedback on the states of the voice recognition.
interface 3 (Figure 3.4)	• An *IntercationMap* shows the mapping of the two multimodal interactions on the top-right corner of the interface. Figure 3.4(a) • By looking at a POI on the interface, the gaze icon and the arrow get blue (*Blue* gaze feedback). Figure 3.4(b) • By touching a POI on the interface, the touch icon and the arrow get blue (*Blue* touch feedback). Figure 3.4(c) • By talking, the speech icon and the arrows get blue (*Blue* speech feedback). Figure 3.4(d) • If a multimodal interaction is detected, the icons and the arrows of the included modalities in that multimodal interactions get green (*Green* feedback). Figure 3.4(e) • This interface also includes *POI size* gaze feedback. Figure 3.2(b)
interface 4 (Figure 3.5)	• Three static icons for each modality are on the top-right corner of the interface to show the supported modalities. (Figure 3.5) • There is no feedback on the states of the voice recognition and no feedback on gaze.
interface 5 (Figure 3.6)	• By talking, a *cursor* gaze feedback appears on the interface. Figure 3.6(a) • By talking, a square blue background (*touch affordance*) appears behind the POIs to afford for touch interaction. Figure 3.6(a) • By looking at a POI or touching it, a *speech bubble* appears on it. Figure 3.6(b)(c)
interface 6 (Figure 3.7)	• On the first try of the interface, the user plays a *coloring game*. The correct color is on each area as a label. Figure 3.7 • The instruction on the game tells users to "Look at the areas or touch them and call the color." • The feel of two multimodal interactions in the *coloring game* are similar to the *base system*. • The tasks in the *coloring game* are 2-step tasks similar to the *base system*. • By touching the areas, a *color palette* immediately appears (MOT).Figure 3.7(b) • By looking at the areas, the shades of their colors change (*Game* gaze feedback). Figure 3.7(a) • If an area is correctly colored, the *color palette*, the *game* gaze feedback, and the label will be removed from it. • By coloring all areas correctly, the game finishes, and the user is directed to the *base system* including an additional *POI size* gaze feedback. Figure 3.2(b) • There is no feedback on the states of the voice recognition.
interface 7 (Figure 3.8)	• If a unimodal interaction is detected, a recommendation for the two multimodal interactions appears on the top of the interface and disappears after 7 seconds. Figure 3.8 • This interface include *POI size* gaze feedback. Figure 3.2(b) • There is no feedback on the states of the voice recognition.

Table 3.1: Details of appearance and behavior of *proposed interfaces*

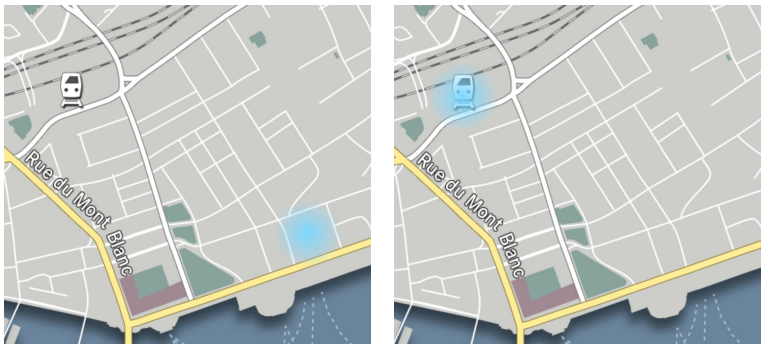

Figure 3.3: *Interface 2*, the appearance of this interface is same as *base system*. It only has additional *cursor* gaze feedback that moves like a cursor on the interface according to eye movements.

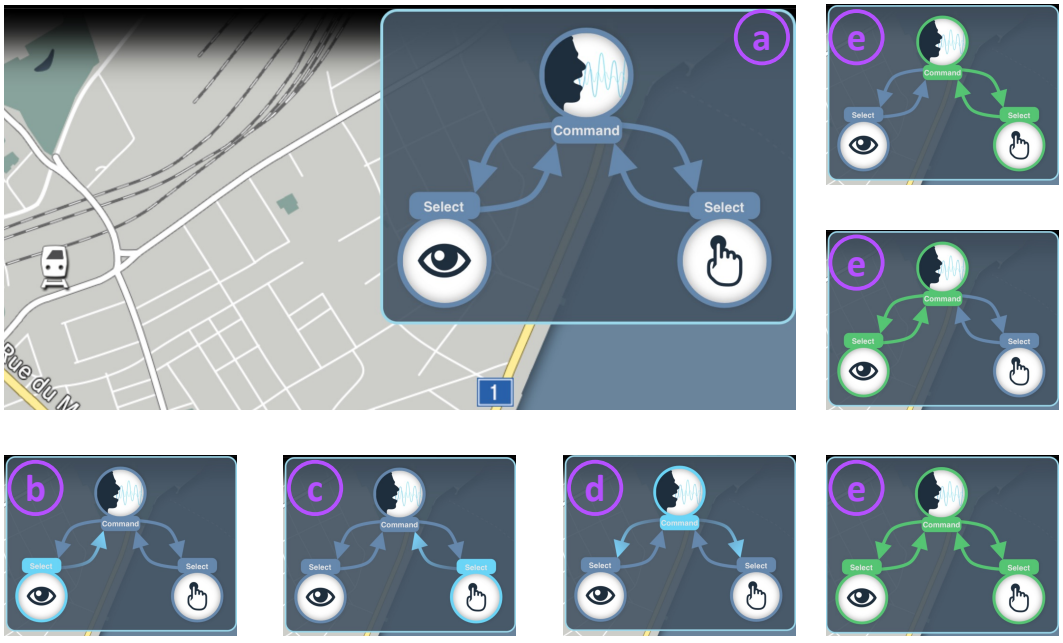

Figure 3.4: *Interface 3*, this interface has the idea of *interface 1* with an addition *InteractionMap* on top-right corner (a). *InteractionMap* includes *blue* feedback on individual modalities (b)(c)(d), and *green* feedback on the detected multimodal interactions (e).

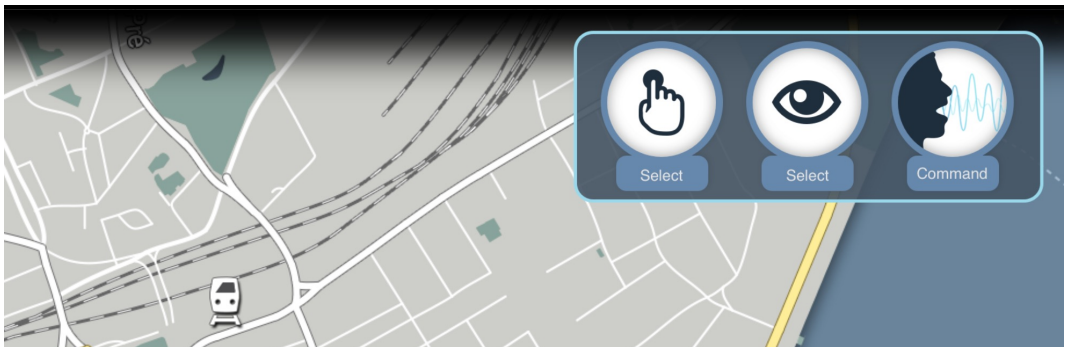

Figure 3.5: *Interface 4*, the appearance of this interface is same as *base system*. It only has additional static icons for each modality on top-right corner of the interface.

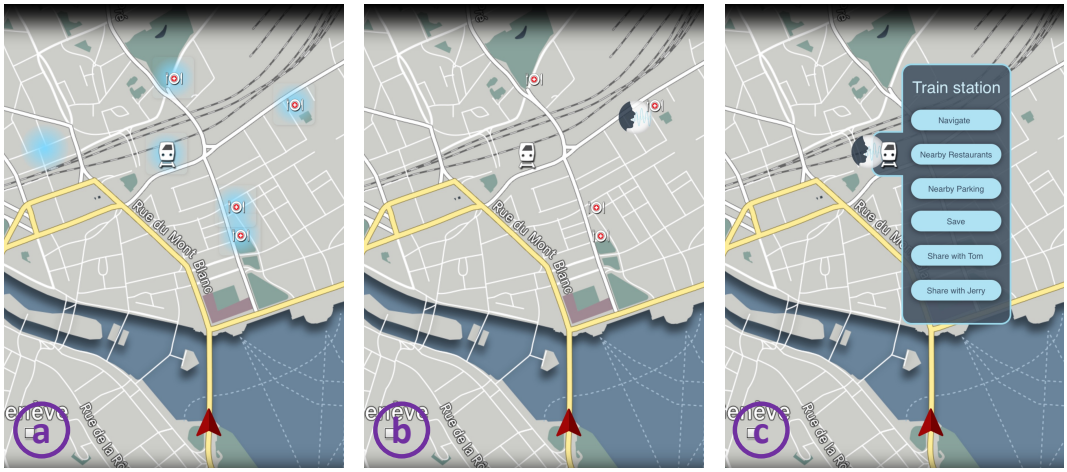

Figure 3.6: *Interface 5* includes *cursor* gaze feedback and *touch affordance* when user is talking (a). When user looks at a POI a *speech bubble* appears on that POI and it disappears when the gaze is off the POI (b). When user touches a POI a *speech bubble* appears on that POI (c) and it disappears when the user close the menu by touching somewhere else on the map.

3.3 Expert Study

We designed an expert study to get a better sense for our *proposed interfaces*. The main goal of this study is to evaluate the discoverability of the *proposed interfaces* to find the two best interfaces as our *candidate interfaces*. However, we were also interested in the whole user experience with the interfaces to make a more sound decision.

Experts helped us to compare the *proposed interfaces* based on discoverability and users experience.

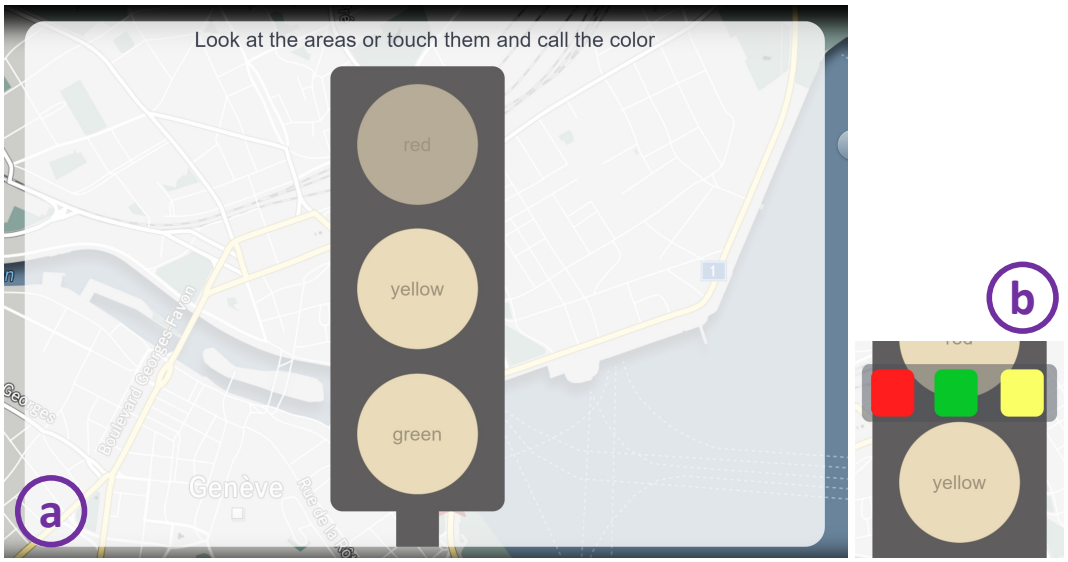

Figure 3.7: *Interface 6* includes *coloring game* at the beginning. The *game* gaze feedback changes the shade of the area's color (a). In this picture, the user's gaze is on the top area. (b) shows the *color palette* that opens on an area when it is touched. When the game is finished, the appearance and behavior of the map application is similar to *interface 1*.

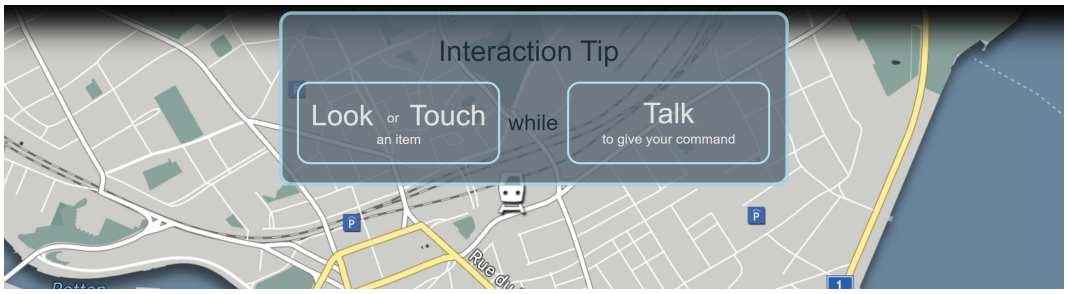

Figure 3.8: *Interface 7* has same appearance and behavior as *interface 1*, except it shows this note when a unimodal interaction is detected.

3.3.1 Heuristics

To reach the goals of this study, we prepared a list of the following heuristics that we evaluated in this study for each of the *proposed interfaces*.

1 Awareness of three modalities: Touch, Gaza, and

Speech.

2 Awareness of two multimodal interactions.

3 Learnability of the multimodal interaction within six trials.

4 Effectiveness of the information on the screen on learning the two multimodal interactions.

5 Effect of usability and complexity of the interface on the general interaction experience.

6 The mental load the interface causes for the users.

3.3.2 Experimental Design

In this study, to compare seven *proposed interfaces*, we designed a within-group study. The order of the interfaces was counterbalanced across eight users using a Latin square design.

3.3.3 Participants

We conducted the study with eight participants, including one pilot (7 male, one female) with an average age of 32.4 years (SD = 7.3 years.) Data from the pilot study are only included in qualitative analyses; however, due to some serious technical issues during the pilot study, some of these data are invalid, and we exclude them. Participants were experts, meaning that they were all familiar with the multimodal interaction concept, and except for one user, they all had experience in user interface design.

3.3.4 Apparatus

For this study, we used a Microsoft Surface Book (312.3mm x 232.1mm x 13.0mm) with its keyboard attached to the touch screen, but we covered the keyboard and touchpad

Figure 3.9: Study setup, including the hardware and the camera that records the hands' interactions.

with white paper with a hole on it for the *space* button. We attached a Tobii Pro Nano to the button edge of the Microsoft Surface for eye tracking. We also used a Sennheiser Headmic calibrated with an M-Audio M-Track to listen to the user's speech input with less noise from the environment. Figure 3.9 shows the study hardware setup.

3.3.5 Tasks

Participants did 6 different location-related tasks to evaluate each *proposed interface*.

In this study, each user did six 2-step location-related tasks for each of the seven *proposed interfaces*. The order of the tasks was fixed across all the interfaces and all the participants. These tasks are defined in the context of a map application as the *base system* is. These tasks in their fixed order are:

Navigation task Navigate to the train station

Restaurant task Find restaurants near the train station

Parking task Find parking near the train station

Save task Save the location of the train station

Tom task Share the location of the train station with Tom

Jerry task Share the location of the train station with Jerry

Throughout the study, an additional *study interface* (Appendix A) did the job of instructing the users through the *proposed interfaces* and the tasks. This interface shows the instructions to the users on a white screen with text on it. There are six different instructions that this interface gives the users. 1) It welcomes the users to the study and asks them to press *space* button to start the first section of the study, which is the first assigned *proposed interface*. 2) It shows the task and asks the users to press *space* button when they are ready. 3) When the users completed the task successfully, it informs them that they have done the task successfully and asks them to press *space* button when they are ready to go to the next task. 4) By successfully finishing the sixth task for a *proposed interface*, it informs users and asks them to evaluate it. 5) When the evaluation of a *proposed interface* is finished, and it is time to go to the next interface, it tells users and asks them to press *space* button to start. 6) When all the *proposed interfaces* are finished, it informs the users that the study is finished.

After seeing a task and pressing the *space* button, users are redirected to the *proposed interface* to do the task. The interface restarts for each task, which means the results of the previous tasks do not remain on the interface. Moreover, when the users finish the task, they are automatically redirected to the appropriate instruction on the *study interface* after 5 seconds.

3.3.6 Study Procedure

After signing the consent form and filling the demographic questionnaire, the study conductor explained the procedure and a guideline for the study. In the guideline, the two

multimodal interactions (*gaze-speech* and *touch-speech*) are introduced. In addition, the guideline includes the heuristics (Section 3.3.1) as the purpose of the study.

After getting enough information about the procedure and the study's goal, participants did a gaze calibration. Users were free to move the display as they wished. The study started while the interaction was recorded on video to facilitate the subsequent analysis. During the study, users were instructed using the *study interface* explained in "Tasks." After finishing six tasks for each interface, users evaluated the interface based on the given heuristics in a questionnaire and then answered some questions in an interview. After finishing all the seven *proposed interfaces*, users ranked them. During the evaluation, users were allowed to look at the interfaces again. Appendix B shows the guideline and questionnaires used in this expert study.

3.3.7 Measurements

Participants evaluated each *proposed interface* one by one in a questionnaire and an interview.

In this study, we gathered both quantitative and qualitative data. After finishing the six tasks for each of the seven *proposed interfaces*, users evaluated the heuristics (Section 3.3.1) with a 5-point Likert scale, and then they took part in an interview.

The second and third heuristics are two attributes (awareness and learnability) of the discoverability as we explained in Section 3.2.1 "Discoverability". In the questionnaire, we asked two separate questions about the awareness of the two multimodal interactions and one question about the learnability of the multimodal interaction. According to Chueke et al. [2017] learning a new interaction within six attempts can be translated to a discoverable interface for that interaction if it is followed by complete awareness. Therefore, we had six tasks for the trial of each interface, and we evaluated the learnability within six tasks. In addition, based on the first heuristic, we asked users to evaluate the awareness of each of the three individual modalities that each of these interfaces provides to get users' opinions on feedback on inputs for each interface.

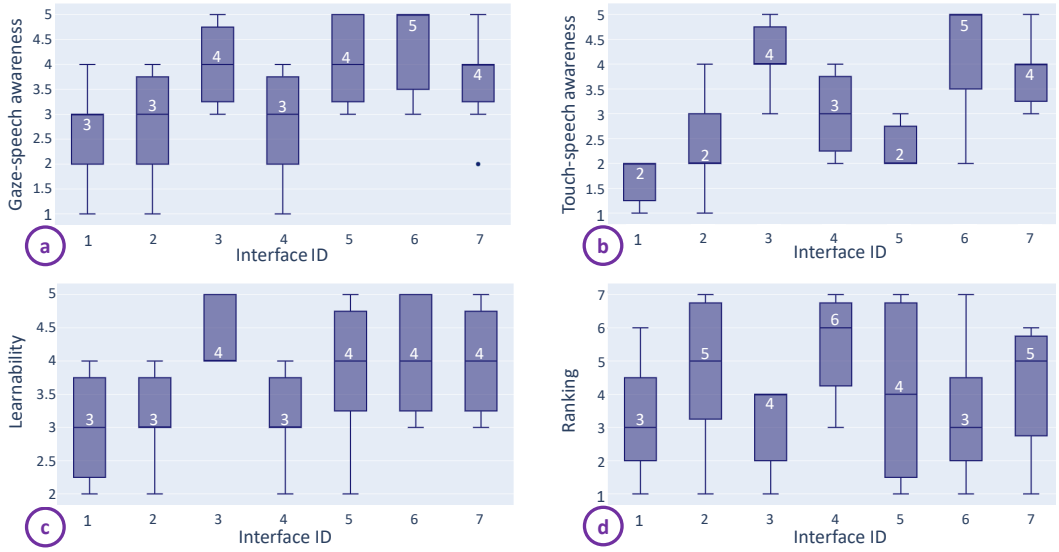

Figure 3.10: Box plots of all scores given to each *proposed interface* for awareness of *gaze-speech* interaction (a), awareness of *touch-speech* interaction (b), and learnability of multimodal Interaction within 6 tasks (c). (d) shows the box plot off the given ranks to each interface. The numbers on the bars show the medians. Whiskers are within a maximum 1.5 IRQ distance.

According to heuristics 4, 5, and 6, we included five questions in the questionnaire to evaluate the effects of these interfaces with the two new interactions and additional design elements for discoverability on users' general experience and the mental load.

Furthermore, we asked users to rank the seven *proposed interfaces* based on their preferences.

Participants ranked the *proposed interfaces* at the end.

3.3.8 Results

Gaze-speech awareness. Figure 3.10(a) shows the box plot of the scores for *gaze-speech awareness* for each of the *proposed interfaces*. As this figure shows, *interface 6* has the best median (median = 5), and followed by *interface 3*, *interface 5*, and *interface 7* (median = 4.) Removing the outlier in this plot does not change the median for *interface 7*.

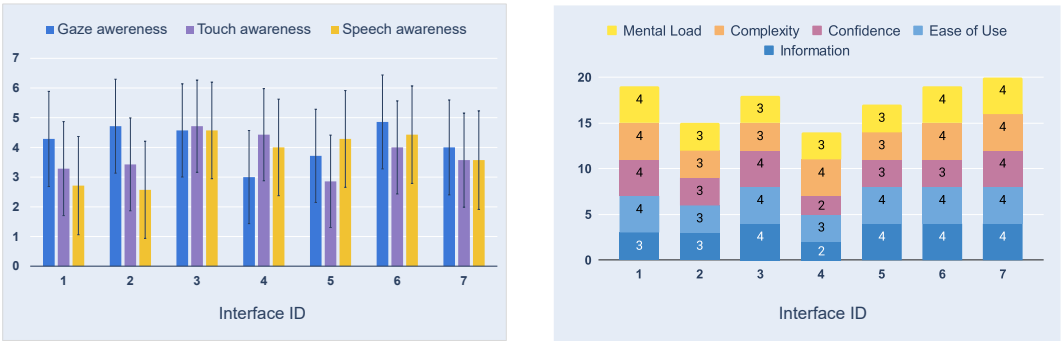

Figure 3.11: The mean score for the awareness level over each individual modality. And bars show the standard deviation (left). The stacked medians of the scores for each usability heuristic (right).

Touch-speech awareness. Figure 3.10(b) shows the box plot of the scores for *touch-speech awareness* for each of the *proposed interfaces*. As this figure shows, *interface 6* has the best median (median = 5), followed by *interface 3* and *interface 7* (median = 4.)

Learnability. Figure 3.10(c) shows the box plot of the scores for *learnability* of the multimodal interactions within 6 tasks for each of the *proposed interfaces*. As this figure shows, *interface 3* , *interface 5*, *interface 6*, and *interface 7* have a higher median (median = 4) than the other interfaces which have median = 3.

Gaze awareness, touch awareness, and speech awareness. Figure 3.11(left) shows the means of these three variables for each of the *proposed interfaces*. As this figure shows, overall performance of *interface 3* and *interface 6* is better than other interfaces based on these three variables.

Ease of use, confidence, complexity, mental load, and information quality. Figure 3.11(right) shows the stacked bars of the median of these variables for each of the *proposed interfaces*. Since the Likert scale of the *complexity* and *mental load* are opposite from the other variables, we reversed the values for these variables in the quantitative evaluation. As

Figure 3.12: Qualitative analysis using post-it notes

this figure shows, the interfaces are mostly similar. However, low medians for *interface 2* and *interface 4* might be an indicator for serious problems with these interfaces.

Ranking. Figure 3.10(d) shows the box plot of ranking. Best median of ranks is for *interface 1* and *interface 6* (median = 3). While, *interface 4* has the worst (median = 6) and it never received the first or second rank. *interface 1*, *interface 3*, and *interface 7* never got the worst rank. The lowest worst rank is for *interface 3*, which is fourth rank.

Qualitative data. Furthermore, in the qualitative data, there were positive and negative comments for each of the *proposed interfaces*. We organized the repetitive comments for each interface on a board using post-it notes (Fig-

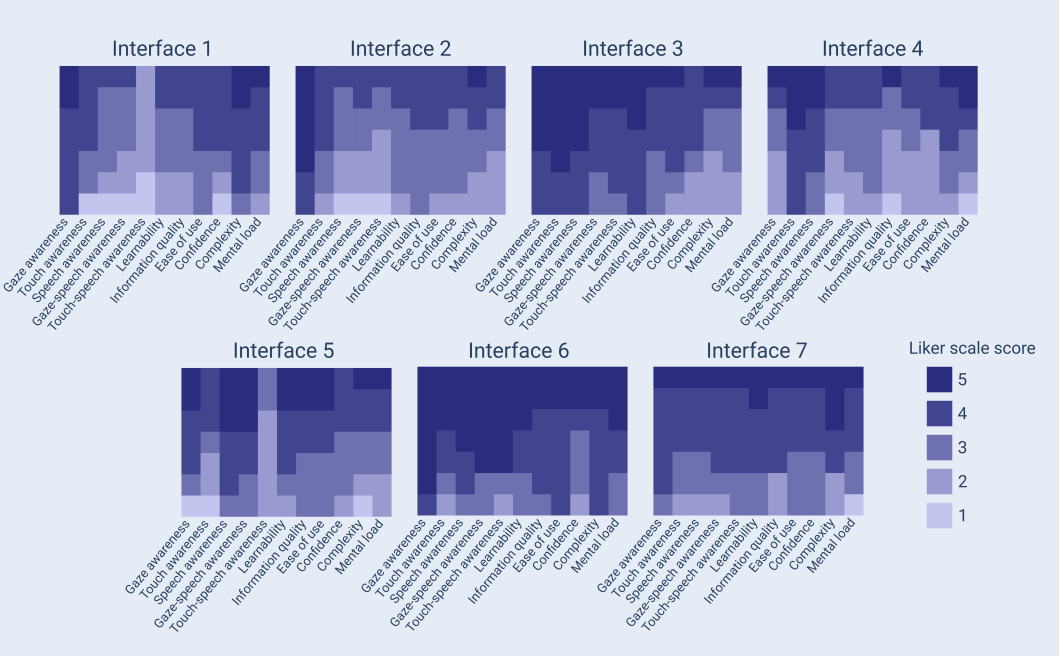

Figure 3.13: Qualitative analysis, this heat map visualize all the answers that all the seven participants (vertical axis) gave for quantitative evaluation of each *proposed interface*.

ure 3.12). In Table 3.2, we have listed the prominent comments for each interface with their number of repetitions.

A the end, we visualized all the quantitative data points in Figure 3.13. This figure shows the sorted scores that participants gave to each variable for each *proposed interface*. The given values for *complexity* and *mental load* are reversed. The darkness of the areas corresponds to scores. The better the score is, the darker the area gets.

3.3.9 Discussion

Quantitative and qualitative data are in favor of *interface 3*, *interface 6*, and *interface 7*.

As Figure 3.13 shows, *interface 3*, *interface 6*, and *interface 7* got the larger dark area in general and also for *gaze-speech awareness*, *touch-speech awareness*, and *learnability* of the multimodal interaction. In addition, the sum of the median of these three variables is highest for the three interfaces (12, 14, 12 respectively.) Comments on these three interfaces

Proposed interface	Repetitive Comments (number of repetitions)
interface 1	• This interface is very good for users who are experts in multimodal interaction.(3) • *POI size* gaze feedback is the favorite gaze feedback of mostly all of the users.(6) • *POI size* gaze feedback should exist in other interfaces.(7) • The behavior of POI size gaze feedback makes users feel they have to look at the POI constantly while giving the speech command.(3) • This interface has serious problems in the discoverability of multimodal interactions.(7) • Lack of discoverable speech interaction.(3)
interface 2	• *Cursor* gaze feedback is not enough and need another gaze feedback like *POI size* to confirm a received gaze on a POI.(7) • *Cursor* gaze feedback is good only at the beginning, and not needed permanently.(4) • This interface has serious problems in discoverability of multimodal interactions.(7) • Lack of discoverable speech interaction.(4)
interface 3	• This interface is good for novice users with the clear *InteractionMap* about the interface capabilities (5), however they might find its design complex.(3) • The structure of *InteractionMap* affords for speech-gaze-touch multimodal interaction.(2) • *Blue* feedback on this interface distracts users.(7)
interface 4	• This interface does not include any gaze feedback, which is necessary for gaze modality in users' opinions.(7) • This interface has serious problems in discoverability of possible interactions.(7) • Lack of feedback on speech input.(2)
interface 5	• This interface is perfect for discoverability of *gaze-speech* interaction, but ruins it for other interactions.(6) • *Touch affordance* and the *cursor* gaze feedback are not understandable.(8) • Lack of feedback on speech input.(3)
interface 6	• This interface provides practice which make it perfect for novices.(4) • There is a problem with transferability of knowledge from *coloring game* to *base system*.(3)
interface 7	• This interface is helpful for novices.(3) • It causes confusion by being difficult to comprehend.(5) • Textual recommendation is not users' favourite.(7)

Table 3.2: Quantitative analysis, repetitive comments for each of the the *proposed interfaces* and their number of repetitions.

also show a higher tendency towards these interfaces. For example, *interface 3* received comments like, P4: "The interaction map informs the user of their capabilities. They should understand how to interact using a multimodal technique within 1 or 2 tasks.", Pilot: "Clear infographic. It is helpful because you see the possibilities.", and P8: "If you plain look at it before you interact just to understand how to interact it is helpful to show that you can combine different modalities.". Comments on *interface 6* were like, P2: "As new users, it shows very well that you can do both. I really like the sentence.", P4: "The game helps the user practice the interactions before executing a task.", and P7: "It can help explaining the user that how you could use gaze, touch and speech together.". Moreover, *interface 7* got positive comments like, P4: "The hint makes sure the user can properly utilize the multimodality.".

Participants prefer interfaces based on visual hints more than textual hints.

In addition, in the interview, participants did some two by two comparisons between some of the *proposed interfaces*. For example, P8 preferred *interface 6* over *interface 7* because it is quicker and needs less mental load. In comparison between *interface 7* and *interface 3*, P5 preferred the first one because it shows hints after a unimodal interaction and does not distract users while interacting. However, P7 and P8 preferred *interface 3* because visual hints are better than textual hints. In addition, in comparison of two visual-based interfaces, *interface 3* and *interface 4*, Pilot and P8 preferred the first one because of multimodal learnability and mature visual feedback, despite distraction, respectively. However, on the other hand, P7 has found *interface 4* clearer.

Feedback on a modality is important to use that modality.

Cursor gaze feedback is incomplete without *POI size* gaze feedback.

On the other hand, participants gave some feedback on the awareness of individual modalities. As the results of *interface 4* show, awareness alone is not enough in order to work confidently with those modalities. Almost all of the participants suggested for including *POI size* gaze feedback and *blue* speech feedback on *interface 4*. Furthermore, they suggested for combining *POI size* gaze feedback with *cursor* gaze feedback on *interface 2*. Participants mentioned that the *cursor* gaze feedback is helpful for the *gaze awareness* at the beginning of the usage, but only if it is combined with *POI size* gaze feedback since it is necessary feedback for the

gaze awareness. Moreover, for some *speech awareness*, participants had additional requirements for more confident use. For example, P4 commented, "I was only aware that I could use speech when looking at a POI and the speech icon came up, some constant feedback that speech was working would help" on *interface 5*, P5 said, "Lighting speech icon up on speech is good thing. It is good to get the idea that did it understand me at all?" on *interface 3*, and P2 stated, "I didn't feel confident while using it." on *interface 4*.

Participants found the *gaze-speech* a more sound multimodal interaction than *touch-speech* for this *base system*.

The other important feedback that participants had, was on *touch-speech* interactions. They believe this interaction in unnecessary on an interface like this. Among four participants who gave this feedback, two of them said gaze is unavoidable when you perform touch, so there is no need for touch when gaze works. And the other two explained the reason as, when you do the touch as part of the *touch-speech* interaction, it would be a lot easier to skip speech and continue the interaction *touch-only*. We can assume two reasons for the later explanation. First, these participants can be "sequential integrators", by Oviatt [2007], in their behavior towards multimodal interaction. Second, the MOT behavior on the *base system* can lead to this mental model over *touch-speech* interaction. As such, we observed fewer *touch-speech* interactions than *gaze-speech* interactions among participants on all the *proposed interfaces*. Specifically on *interface 6*, even thought the *coloring game* was for both interactions, most of the participants did not even try *touch-speech*.

To conclude, data leans towards *interface 3*, *interface 6*, and *interface 7*, but the fact that no interface got a median = 5 for *learnability* shows there are some problems with all of the interfaces including these three. Table 3.2 shows the most prominent problems with each of the interfaces. However, having three interfaces as our *candidate interfaces* for the Evaluation Study would exceed the time frame of this thesis. Therefore, we exclude *interface 7* since the problems with this interface would take more time to solve than what is available. Also, its performance was not as good as the other two interfaces.

Based on these findings, we picked *interface 3* and *interface 6* as our *candidate interfaces* to proceed to Chapter 4 "A Com-

parison of Different Interface Designs for Better Discoverability" after making the necessary changes we explain in the following. In addition, from now on, we refer to these interfaces with a clear name. Therefore, we use name *InteractionMap* instead of *interface 3* and *Game* instead of *interface 6*.

Interface 3 and *interface 6* are picked as the *candidate interfaces*, with *InteractionMap* and *Game* as their new names, respectively.

3.3.10 Candidate Interfaces Appearances and Behaviors

Based on the results from expert study, we applied some changes to *InteractionMap* and *Game* to hopefully solve the important issues in these interfaces. As Table 3.2 shows, the main problem with *InteractionMap* is complexity and distraction, and with *Game* is knowledge transferability. Table 3.3 lists the design details of these two interfaces including the applied changes.

Candidate Interface	Details of Design
InteractionMap (Figure 3.14)	• A new *IntercationMap* shows the mapping of the two multimodal interactions on the top-right corner of the interface (Figure 3.14(a)). To remove complexity, we replaced arrows in the old design with a plus icon, and separated the mapping of the *gaze-speech* and *touch-speech* interactions. • By looking at a POI on the interface, the gaze and plus icons get blue (*Blue* gaze feedback), but to reduce distraction, this feedback remains for 2 seconds. Also this interface includes *POI size* gaze feedback. Figure 3.14(b) • By touching a POI on the interface, the touch and plus icons get blue (*Blue* touch feedback). Figure 3.14(c) • By talking, the speech and plus icons get blue (*Blue* speech feedback). Figure 3.14(d) • If a multimodal interaction is detected, the background of the detected multimodal Interaction gets green (*Green* feedback). However, in the new design, if both multimodal interactions are detected, only the back ground of *touch-speech* gets green. Figure 3.14(e)
Game (Figure 3.15)	• On the first try of the interface, the user plays a *coloring game*. The correct color is on each area as a label. Figure 3.15(a)(b) • To solve the knowledge transferability problem, the purpose of the game is explained at the beginning. Figure 3.15(c) • To make sure that users practice both *gaze-speech* and *touch-speech* interactions, we separated the *coloring game* of these two interactions. • The instructions on the game include examples [Srinivasan et al., 2019]: "Look at the areas and say for example "color it red"" for *gaze-speech*, and "Touch the areas and say for example "color it red"" for *touch-speech*. • The feel of two multimodal interactions in the *coloring game* are similar to the *base system*. • The tasks in the *coloring game* are 2-step tasks similar to the *base system*. • By touching the areas, a *color palette* immediately appears (MOT). Figure 3.15(a) • By looking at the areas, the shades of their colors change. (*Game* gaze feedback) Figure 3.15(b) • If an area is correctly colored, the *color palette*, the *game* gaze feedback, and the label will be removed from it. • By finishing the *coloring game* for both interactions, the user is directed to the *base system* including an additional *POI size* gaze feedback and *speech wave* for speech feedback to improve *speech awareness*.

Table 3.3: Appearance and behavior of *candidate interfaces* after changes based of the results from the expert evaluation.

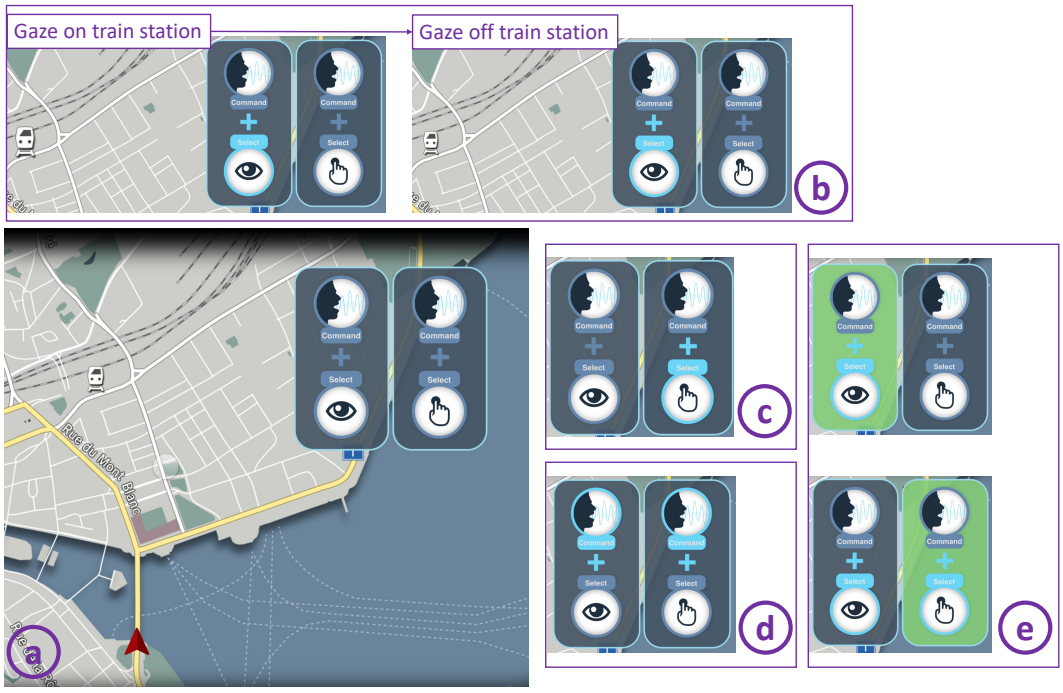

Figure 3.14: The changes in *InteractionMap* after the expert evaluation include: changes in the design of the *InteractionMap* (a), 2-second delay on disappearing the *blue* gaze feedback (b), and clearly separating the *green* feedback on multimodal interactions (e). However, the *blue* touch and speech feedback have similar behavior as old interface (c)(d).

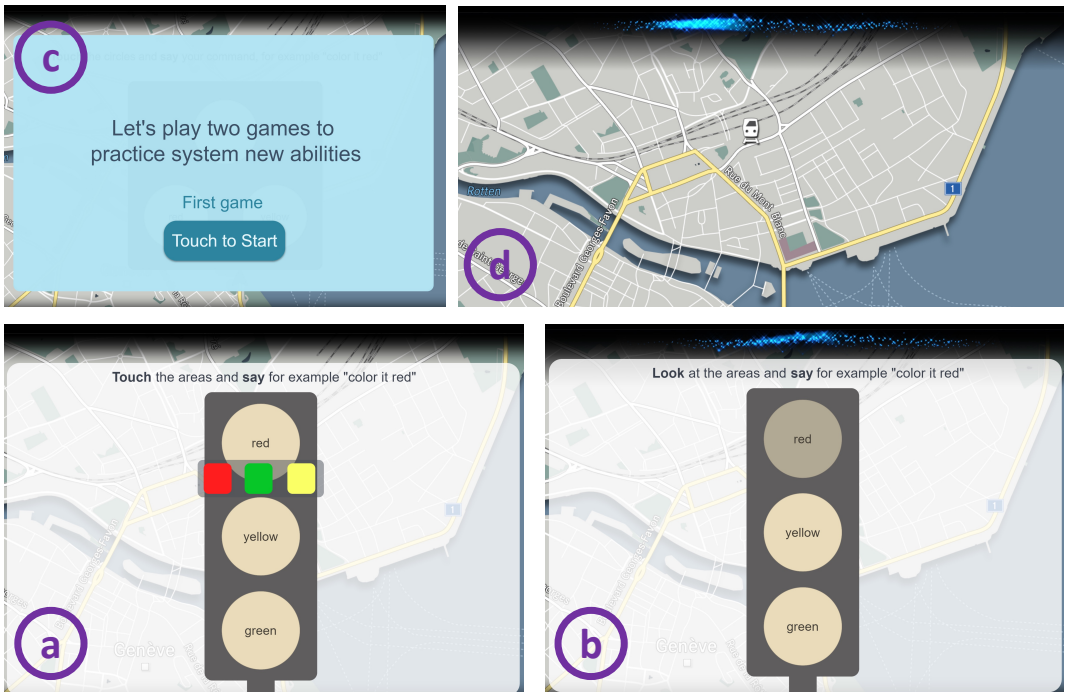

Figure 3.15: After the expert evaluation, *Game* includes separated *coloring games* for each multimodal interactions (a)(b), includes a description for the purpose of the game at the beginning (c), and a *speech wave* on both map application and game (b)(d).

Chapter 4

A Comparison of Different Interface Designs for Better Discoverability

In this chapter, we describe the user study we conducted to compare the two *candidate interfaces* (*InteractionMap* and *Game*) in Section 3.3.10 "Candidate Interfaces Appearances and Behaviors" with a *Baseline interface* for discoverability of *gaze-speech* and *touch-speech* multimodal interactions .

We compare *Game* and *InteractionMap* with *Baseline*.

4.1 Baseline Interface

According to one of the motivations behind this thesis, the discoverability of multimodal interactions on an interface is an open question in the literature. However, in the studies in the multimodal area, conductors make users aware of the multimodal capability of the system as part of the study procedure [Srinivasan et al., 2020, 2019, Oviatt et al., 1997]. Therefore, we decided to do the baseline by introducing the new multimodal interaction techniques at the beginning of the study. To do so, we created a video . The video contains information on available interaction possibilities to increase

Baseline used a video to introduce *gaze-speech* and *touch-speech*.

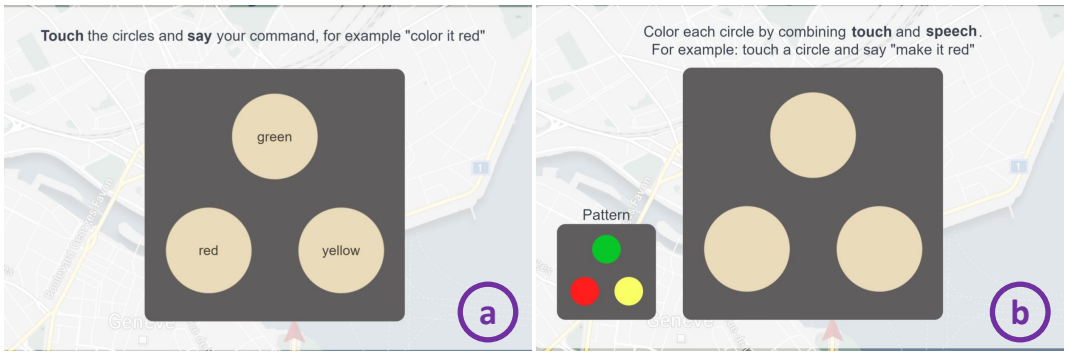

Figure 4.1: *Game* interface, we changed the design of the *coloring game* after two rounds of pilot study. (a) and (b) show the changes after the first and second round respectively for *touch-speech* part of the game.

awareness and examples to teach how to perform the possible multimodal interactions to increase learnability. The awareness part of the video includes two sentences each for one of the possible interaction techniques: "You can combine your gaze with your speech" for *gaze and speech* interaction, and "You can combine your touch with your speech" for *touch and speech* interaction. The example part of the video is followed by each awareness sentence showing a person doing the *navigation* task to the train station. In addition, as Freeman et al. [2009] used short videos ($\leq$ 3.5 seconds) as their base condition for teaching hand gestures, we also kept our baseline video short.

4.2 Pilot Study

We did a pilot study for the evaluation study, which is described in Section 4.3 "Evaluation Study". Six persons participated in this pilot study (male: 5, female: 1) with an average age of 26.2 years (SD = 3.2 years.) Based on the results of this pilot study, we made some changes in the interfaces.

4.2.1 Changes in Interface Designs

The pilot study revealed that the instruction in the *coloring game* is difficult to understand.

In the *Game* interface, users had problem with correctly understanding the *coloring game*. Instead of focusing on how to do the coloring task, they were struggling to understand what the task is. We thought about three possible reasons for this problem. First, the problem could be the unclear instructions that the game gave the users to increase their awareness and teach them the *gaze-speech* and *touch-speech* interactions (Figure 3.15(a)(b)). Second, the color label on each circle, which is unnatural in a *coloring game* context could cause this problem. And third, we thought that a traffic light might be a misleading object to color, because users might think that the goal of the game is to completely color the traffic light by getting the last circle green as a pass to go to the next section. However, the goal of the game is to use the familiar coloring task and practice the system capability without paying attention to what they are coloring.

Within two rounds we found the final look and instruction for the *coloring game*.

To solve the problem with the *Game* interface, we first changed the instruction in the *gaze-speech* game from "Look at the areas and say for example "color it red"" to "Look at the circles and say your command, for example "color it red"", changed the instruction in the *touch-speech* game in the same way, changed the traffic light to three circles on the corner of a triangle, and we did not change the color labels (Figure 4.1(a)). Two users tried this solution and the results did not improve. Therefore, in the the next step, we changed the instruction in the *gaze-speech* game to "Color each circle by combining look and speech. For example: look at a circle and say "make it red"", changed the instruction in the *touch-speech* game in the same way, we kept the three circles, but we removed the labels and added a pattern of the correct coloring on the corner (Figure 4.1(b)). Again two users tried the new solution and the observation showed improvement in users' confidence when doing the game.

Based on the pilot study, we added *cursor* gaze feedback.

The other problem we observed in the pilot study was that the *POI size* gaze feedback on the interfaces is not enough especially in the cases that the gaze calibration is not perfect or the users are not familiar with the gaze modality.

Figure 4.2: New Design of *cursor* gaze feedback. This gaze feedback and *POI size* gaze feedback exist on the interface at the same time (right).

On the other hand, according to the expert study, experts named the *cursor* gaze feedback as a helpful feedback for the novice users. Therefore, we decided to add the *cursor* gaze feedback to *candidate interfaces* as well as *Baseline interface*. However, because of the occlusion problem that two participants named for the old design in expert study, we designed a new *cursor* gaze feedback (Figure 4.2).

4.3 Evaluation Study

The goal of this study is to measure the discoverability of the two multimodal interactions (*gaze-speech* and *touch-speech*) on the three *interface conditions* (*Game*, *Interaction-Map*, and *Baseline*). To design this study, works of Chueke [2016], Goguey et al. [2018], Walter et al. [2013] on discoverability of gestural interactions inspired us.

4.3.1 Experimental Design

Since our goal is highly susceptible to the learning effect, we designed a between-group study. Therefore, we randomly assigned each user one of the *interface conditions* with keeping the total number of users for each condition equal.

Each user did 12 tasks that included a repeat of six different

tasks. The order of the tasks was counterbalanced across the users using a Latin square design with the condition of not having a repetitive task in the first six tasks and not having two same tasks in a row.

In the *Game* condition, the first game that user did could be either *gaze-speech* game, or *touch-speech* game. We kept the starting game random across the users of this condition.

4.3.2 Participants

We conducted the study with 36 participants who did not participate in the expert study. The participants were 20 to 60 years old with an average of 36.98 years (SD = 9.9 years.) 30 participants were male, 13 were female. Participation in this study had no precondition.

4.3.3 Apparatus

In this study, we used the same hardware as in the expert study (Section 3.3.4 "Apparatus".) We used a Microsoft Surface Book (312.3mm x 232.1mm x 13.0mm) with the touch screen attached to its keyboard, which was covered with white paper. The Tobii Pro Nano was attached to the bottom edge of the Microsoft Surface to track eye movement. Also, the speech input was captured by a Sennheiser Headmic calibrated with an M-Audio M-Track.

4.3.4 Tasks

Each user did 12 location-related tasks on the assigned *interface condition*.

In this study, users did 12 tasks in the *interface condition* that was assigned to them. These 12 tasks included six different tasks, each of them two times. These six tasks are the same 2-step location-related tasks as in the expert study (Section 3.3.5 "Tasks".) In addition, we also used the same *study interface* used in the expert study, except this time it was 12 tasks for only one *interface condition*. This *study interface* gives the users the tasks one by one, and after giving

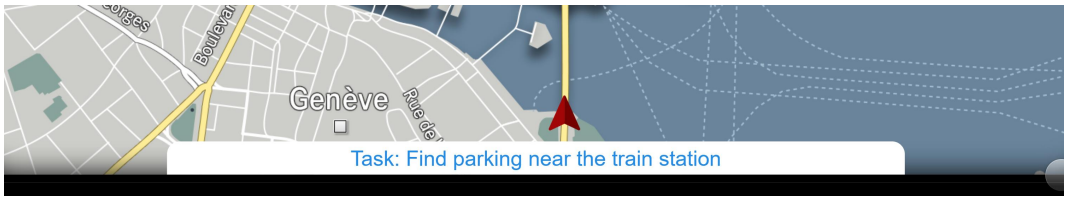

Figure 4.3: Tasks are shown to the users as a subtitle on the interface.

each task, it redirects the users to the restarted state of the *interface condition*. However, according to our observations in the expert study, when the *study interface* first gave the task at each step and then redirected the users to the *interface condition*, users sometimes forgot the task. Therefore, in addition to showing the task before redirecting users to the *interface condition*, we added the task as a subtitle to this interface (Figure 4.3).

Task was shown on the interface like a subtitle.

When the task is finished, user can go to the next task by pressing *space* button.

When the task is finished, the subtitle changes to "Task finished. Press space." By pressing the *space* button, the users are redirected to that instruction of *study interface* which tells users, "You have successfully completed the task. Press space to go to the next task." The other five tasks are also possible when users are doing a task, but users should finish the assigned task at each step to proceed to the next task.

4.3.5 Study Procedure

This study started with a short instruction about the study. Same as Goguey et al. [2018] and Walter et al. [2013] works on discoverability, this instruction did not include any information on the possible interactions with the system and also supported modalities of the system. However, users wore a microphone and did a gaze calibration right at the beginning. Therefore, to be fair to touch modality, if users struggled to tap a button or a POI on these interfaces, the study conductor made them aware of the touch screen. The distance and the angel of the display could change based on the gaze calibration and the users' preferences.

The information we gave to the users as the instruction con-

tained: information on the number of tasks, information on what the tasks are, information on the *study interface* that guided them through the task that they should do in each step, and the goal of the study which was stated as "accomplish the task in a natural speed." What we observed in the "Pilot Study" was that users stuck to the first interaction technique that works for the first task, and that was *speech-only* and *touch-only* for half of the users. Also, as one of the users stated: "Next time, I would explore the interface more instead of focusing on the tasks and on the time I spend finishing them (I wasn't in a hurry, just too focused on them.)", we observed that users rush into doing the tasks without exploring the system. Our conclusion based on the observations was that the instruction is probably the problem. The fact that the instruction has too much focus on the tasks without giving users a correct overview of the study's goal causes the observed behaviors. As a result, we changed the instruction to a more realistic goal: "In this study, you will try a new system. To do that, you will be given 12 exemplary tasks." However, later on, as we will discuss in Section 4.3.10 "Limitations", we found that there were also other probable reasons behind this behavior.

Given study instruction at the beginning changed after the pilot study.

After the instruction and setup, a video recording of users' hands interacting with the interface started. Then users started the tasks using the *study interface*. In the *Game* and *Baseline* conditions, users played the game or watched the video after seeing the first task. When the tasks were finished, like Goguey et al. [2018] work, the conductor asked the user to describe the possible interactions on their own word, and then they filled a questionnaire. In the end, users participated in a short interview.

The other observation we made in the "Pilot Study" was that the users are not comfortable with speech. One of the users stated: "I saw the wave icon on the screen in the beginning. So I should have tried speech out, but I was too shy/unsure if it works. I didn't want to disturb anyone." The study setup was first in an area where there were other people, which could have made the user uncomfortable. Therefore, we decided to move the setup to a private location.

4.3.6 Measurements

We measured *discoverability*, *awareness*, and *learnability* of the two multimodal interactions.

This study is designed to measure the *discoverability* of *gaze-speech* and *touch-speech* multimodal interactions in the three *interface conditions*. As we mentioned in Section 3.2.1 "Discoverability", two key attributes of discoverability are *awareness* and *learnability*.

Based on the work of Chueke et al. [2017], a discoverable gestural interaction should be mindfully learnable within the first six attempts. To measure discoverability, they introduce a formula. We recreated their formula for our study. The following shows the formula we used to measure the *discoverability* of *gaze-speech* and *touch-speech* for each user:

Discoverable If the user learns to perform the interaction within six tasks and is able to describe the interaction.

Partially discoverable If the user learns to perform the interaction from the seventh task onward and is able to describe the interaction.

Failure If the user does not learn to perform the interaction or is not able to describe the interaction.

Two methods are used for measuring *gaze-speech* and *touch-speech awareness*.

We did video recordings of the interactions with the system and captured log files to label each task in a study with an interaction technique and used these labels for measuring *discoverability*. Moreover, in measuring *learnability* in one study, if there is at least one task among 12 tasks with *gaze-speech* or *touch-speech* label, that multimodal interaction is learnable in that study; otherwise, it is not. Furthermore, we measured *awareness* using two methods. First, after finishing the tasks, we asked the awareness question: "What are the possible ways to do a task in this interface? Please describe all of them in your own word." If *gaze-speech* or *touch-speech* is described in the answer, the user is aware of that multimodal interaction based on the awareness question. Second, without giving any extra information on multimodal interactions, we asked users to fill a questionnaire.

In this questionnaire, we asked users two questions to evaluate their *awareness* of *gaze-speech* and *touch-speech* multimodal interactions in a 5-point Likert scale. We used the results from the first method in the formula for *discoverability*. In addition, the questionnaire included three Likert scale questions to evaluate the *gaze awareness*, *touch awareness*, and *speech awareness*. In the end, we asked participants if they were familiar with the multimodal interaction concept before the study. Appendix C shows the questionnaire.

At the end, user were asked about their prior multimodal knowledge.

4.3.7 Labelling of Data

Total number of finished tasks in the whole study is 436 (36 participants $\times$ 12 tasks), and we call each interaction attempts, for finishing a task, an interaction trial.

To label each interaction trial with the intended interaction technique (*gaze-speech*, *touch-speech*, *touch-only*, or *speech-only*), we followed some rules. If the speech command includes "(the) train station," the interaction type is *speech-only* unless users say a clear statement in the post-study interviews that they intended for *gaze-speech* or *touch-speech*. The latter situation occured in 25 trials across four participants, for example one of them said in the interview "I guess you can also use only speech, I didn't try it.". In addition, if the "train station" is used with definite articles "this" and "that", and the log file shows an occurrence of *gaze-speech* or *touch-speech*, the interaction type is one of the multimodal interactions. However, no speech command in *touch-speech* trials included "train station." If the log file shows both *gaze-speech* and *touch-speech* for a trial, the *touch-speech* is selected as the intended interaction technique.

We followed some rules to label each interaction trial with the intended interaction technique.

Moreover, participants did not necessarily finished a task within one interaction trial. 52 tasks were finished by more than one unique interaction trial. In 12 cases, participants did another trial because their previous trial was not successful due to the technical issues. In 4 cases, they did a new trial because they did not realize that the task is finished. And in 39 cases, at least one of the trials was invalid.

In some cases, users did more than one interaction trial to finish a task.

Not all the interaction trials were valid.

There are five groups of invalid interaction trials. 1) Interactions that are correctly multimodal, but they do not lead to finishing the task, because they are not intended for the given task. For example, doing a *navigation* task instead of a *parking* task. These invalid interaction trials occured in 12 out of 436 tasks. 10 of them were *navigation* task instead of the given task in the *Baseline* condition, and among these 10 tasks, seven of them were the first given task. 2) The second type of invalid trials happened for the *parking* and *restaurant* tasks. When these tasks are finished, some new POIs for parking and restaurants appear on the map. In 15 tasks, users performed multimodal interactions (86.6% *gaze-speech* and 13.3% *touch-speech*) to navigate to one of the new POIs. Although this interaction is a correct multimodal interaction, we count it as an invalid trial since we aimed for doing the measurements for exactly the given task, and our prototype did not support it. 3) The third group of invalid trials are those that includes unsupported speech commands in a *gaze-speech* multimodal interaction. These speech commands in these trials were short commands like: "select", "press", "confirm", and "options". Six participants performed this invalid interaction trial within 10 tasks. 4) The other invalid trial happened only three times for either *Tom* or *Jerry* tasks. For example one participant performed it like this: "Hey Jerry, this is the location of the train station." while she was looking at the train station. 5) The last group of invalid interaction trials are those that include incomplete speech commands in *Tom* or *Jerry* tasks. for example a participant said: "Share location." An interaction like this happened only in two tasks. The forth and fifth groups are counted as invalid because the *base system* could not support them, however, they were actually multimodal.

We labeled each task with the last valid interaction trial for that task.

We labeled each of the 436 tasks with only one valid interaction technique. We ignored the invalid trials, and among valid trials, we picked the last valid trial that the user performed. We could also pick the first valid trial, since the effect is only marginal (one participant in *Baseline*, and one in *InteractionMap* conditions.)

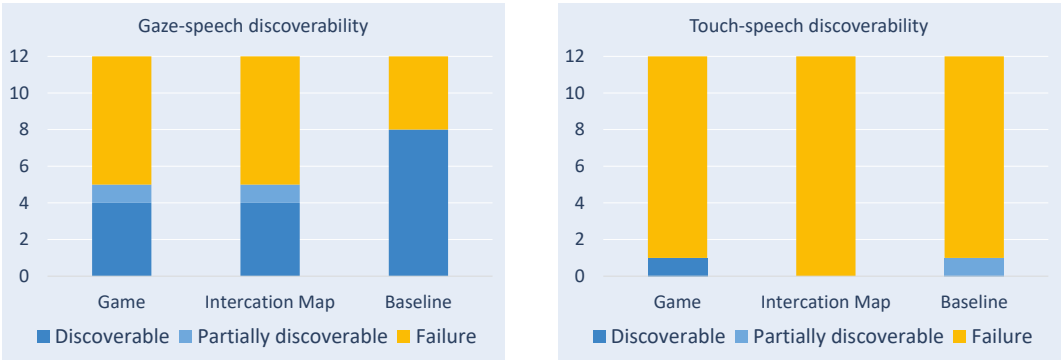

Figure 4.4: Number of discoverable, partially discoverable, and failure cases for *gaze-speech* interaction (left) and *touch-speech* interaction (right) in three *interface conditions*

4.3.8 Results

Gaze-speech discoverability and touch-speech discoverability. Figure 4.4 shows the number of *discoverability* cases for each of the *interface conditions*. A Kruskal Wallis test shows no significant difference between *interface conditions* in *discoverability* of both *gaze-speech* and *touch-speech*.

Gaze-speech learnability and touch-speech learnability. Figure 4.5(left) shows the number of cases for each *interface condition* that includes at least one *gaze-speech* interaction within 12 tasks, and Figure 4.5(right) shows it for *touch-speech*. A Kruskal Wallis test shows no significant difference in *learnability* of *gaze-speech* and also *touch-speech* between all three *interface conditions*.

Gaze-speech awareness and touch-speech awareness. Figure 4.6(left) shows the number of participants who described *gaze-speech* in answering the awareness question, and Figure 4.6(right) shows it for *touch-speech*. A Kruskal Wallis test revealed a significant effect of *interface condition* on *awareness* of *touch-speech* ($\chi^2(2)=10.0$, $p < 0.006$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between *Interaction-Map* and *Game* ($p < 0.007$, $r = 0.63$).

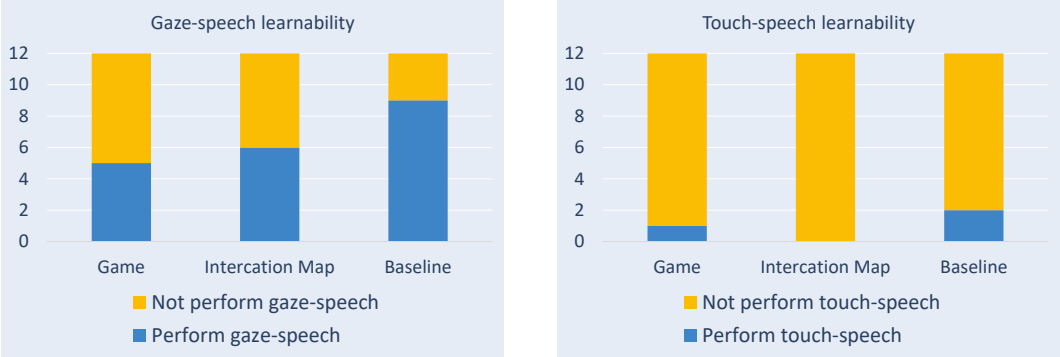

Figure 4.5: Number of cases that performed (learned) *gaze-speech* interaction (left) and *touch-speech* interaction (right) at least once within 12 tasks in each *interface condition*

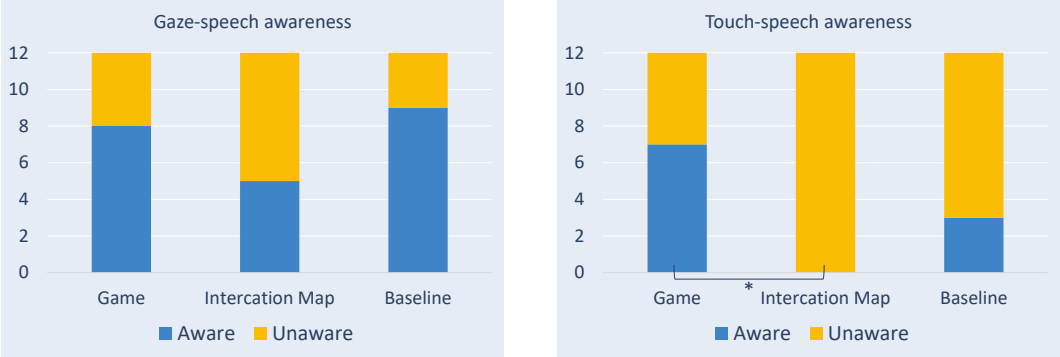

Figure 4.6: Number of cases that were aware of *gaze-speech* interaction (left) and *touch-speech* interaction (right) in three *interface conditions* based on the awareness question

However, results from the questionnaire, the second method of measuring *awareness*, shows no significant difference between the three *interface conditions* in both multimodal interactions using a Kruskal Wallis test. Figure 4.7 shows the box plots of these measurements.

Gaze awareness, touch awareness, and speech awareness. The box plots of these three variables are shown in Figure 4.8. There is no significant difference for *gaze awareness* between *interface conditions*. However, a Kruskal Wal-

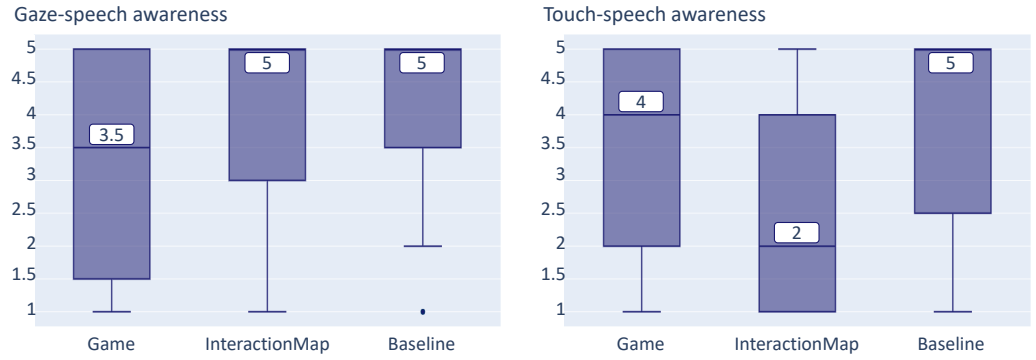

Figure 4.7: Box plots of *awareness* of *gaze-speech* interaction (left) and *touch-speech* interaction (right) in three *interface conditions* based on the questionnaire. Numbers on the box plots show the median.

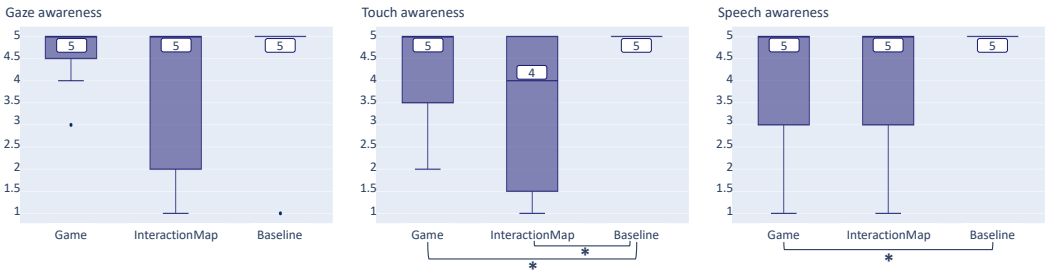

Figure 4.8: Box plots of awareness of each modality in three *interface conditions*. The numbers on the box plots show the median.

lis test revealed a significant effect of *interface condition* on *touch awareness* ($\chi^2(2)=8.3$, $p < 0.02$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between *Game* and *Baseline* ($p < 0.04$, $r = 0.50$) and between *InteractionMap* and *Baseline* ($p < 0.01$, $r = 0.56$). Furthermore, same test showed a significant effect of *interface condition* on *speech awareness* ($\chi^2(2)=5.8$, $p <0.05$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between *Game* and *Baseline* ($p < 0.03$, $r = 0.50$).

Familiarity with multimodal interaction concept. 58.3%, 33.3%, and 58.3% of participants stated that they had prior knowledge of multimodal interaction in *Baseline*, *Game*, and

InteractionMap conditions, respectively.

4.3.9 Discussion

As the results show, the *Game* and *InteractionMap* conditions are not significantly different from *Baseline* condition in *discoverability*. However, our prediction for this study was a high *discoverability* of *gaze-speech* and *touch-speech* in *Baseline* condition, the results show a poor performance of this condition in *discoverability* of *touch-speech*. Since *awareness* for *touch-speech* is rather high (Median = 5) in this condition (Figure 4.7(right)), the investigations continued with the probable issues in *learnability* of *touch-speech*.

Baseline did not perform as expected in *touch-speech discoverability*.

The qualitative data shows that there are two main reasons for the failure of *Baseline* condition in *discoverability* of *touch-speech* multimodal interaction. First, technical issues can cause damage in *learnability* and *awareness*, and as a result, *discoverability*. Second, the fact that there is another new multimodal interaction which has the support of the setup, and has a more exotic modality, gaze, involved with an explicit *cursor* gaze feedback, can result in neglecting *touch-speech* and therefore a bad *discoverability*.

A detailed analysis shows that among 12 users in *Baseline*, five users tried to perform *touch-speech* in their interaction trials and only one case resulted in *discoverability* of *touch-speech*. In two of the cases, a technical problem while performing *touch-speech* was reason for failure in both *learnability* and *awareness* of this interaction. For example, one of the participants commented on the low *touch-speech awareness*: "Did not work the way I thought it should." In the other two cases, participants tried to perform *touch-speech*. However, they explained in the interview that their intention of doing touch is to see the possible options and probably not to perform *touch-speech*. Therefore, we continued the investigations by analyzing the effect of *discoverability* of *gaze-speech* on *learnability* of *touch-speech*. As Figure 4.4 shows, *gaze-speech* had shown a rather high *discoverability* in *Baseline*. The four failures of *discoverability* of *gaze-speech* were also failures of *touch-speech*. The failure reason was video

in three cases. The problems were no clear instruction (2 cases), no clear way of how-to confirm gaze (2 cases), language issue (2 cases), and lab-condition anxiety (2 cases).

gaze-speech interaction affected *touch-speech discoverability*.

However, among eight cases that discovered *gaze-speech* in *Baseline* condition, five of them were aware of *touch-speech* (based on questionnaire). Investigation on the behavior of those participants who tried *gaze-speech* shows that most of them had a high focus on *gaze-speech* interaction because of three reasons. First, the video did not contain enough information for them to learn *gaze-speech*, so they had to spend time to learn it mostly themselves (2 cases). Second, *gaze-speech* was comfortable and easy in this study setup, so there was no need to try the other interaction (2 cases). Third, experience with gaze made them curious to try it in this system (2 cases).

Some changes can improve the video in *Baseline*.

Therefore, we can suggest some improvements in the video. The video should be longer and include more detailed information. For example, multimodal interaction concept should be introduced including information on the role of each modality and how to perform each of them in the multimodal interactions. Moreover, the high frequency of the first type of invalid interactions in the *Baseline* condition (in Section 4.3.8 "Results"), shows that video needs practice. Therefore, including more than one example can replace this need for practice and help to make the interaction clearer.

The other interesting result that is worth investigating is *discoverability* in *InteractionMap* condition. In this condition, 56% discovered *gaze-speech*, but no one discovered *touch-speech*. To find out the influential factors in this result, we started from the analysis of the *awareness*. As Figure 4.7 shows, there is a big difference between *awareness* for *gaze-speech*, and that for *touch-speech* in this condition. Since *InteractionMap* tries to show both multimodal interactions in similar ways, we dig deep into qualitative data from interviews to find the reasons. The qualitative data shows that reasons for high *awareness* for *gaze-speech* are prior knowledge of multimodal interaction (7 cases), and therefore, a quick notice was enough for them (4 cases) to realize the possibility of multimodal interaction and discover it. How-

Gaze-speech awareness in *InteractionMap* depended highly on familiarity with multimodal concept.

ever, the prior knowledge did not improve *awareness* for *touch-speech*. The low *awareness* for *touch-speech* is because participants have already found a working interaction technique (6 cases), and the study setup does not afford for touch (5 cases).

InteractionMap had complexity issues for novices.

Therefore, a lot of information on *InteractionMap* was unnecessarily extra for users who were familiar with the multimodal concept, so they overlooked the details of it. However, that information was few or too vague for the unfamiliar participants with the multimodal interaction concept or those who do not have experience in such interaction, especially on how-to perform gaze and *gaze-speech*. In addition, study did not invite for exploring new things. As such, they tried to ignore the *InteractionMap* after finding a working interaction technique.

The *coloring game* improved awareness of both multimodal interactions.

In the *Game* condition, more participants described *touch-speech* in answering the awareness question compared to the other *interface conditions* (Figure 4.6(right)). The high *awareness* was caused by the *coloring game* in most cases (5 cases). However, there are three reasons for poor *touch-speech learnability* in this condition. First, touch can be replaced by gaze if it works properly (4 cases). Second, participants had a problem finding the right speech commands in the map application (3 cases). Last but not least, participants could not apply their knowledge from the *coloring game* to the map application (5 cases). The last two reasons, affected *learnability* of *gaze-speech* as well. The lack of knowledge transferability between *coloring game* and the map application was harmful to *learnability*; however, one participant commented on an abstract idea for the *coloring game* as an introduction: "Maybe an abstract introduction is not a bad idea, because if you have different use cases, a specific introduction would not fit to another use case. But if there is this one use case, it would be good idea to have an introduction more about navigating on the map."

Game had knowledge transferability issues.

For novices, *Game* performed better than *InteractionMap* in *discoverability* of *gaze-speech*.

Furthermore, we did a separate analysis of the three *interface conditions* for only those cases in which participants were unfamiliar with multimodal interaction concept before the study. In *Baseline*, *Game*, and *InteractionMap* conditions, 60%, 25%, and 20% of the cases successfully dis-

covered *gaze-speech* respectively. However, there is no successful *touch-speech discoverability* recorded for these three conditions when participants had no prior knowledge of multimodal interaction.

In summary, *InteractionMap* and *Game* performed similar to *Baseline* in *discoverability* of *gaze-speech* and *touch-speech*. *Game* could increase the *touch-speech awareness* better than the two other *interface conditions*, and *gaze-speech awareness* almost similar to *Baseline*. Moreover, it performed better than *InteractionMap* in *gaze-speech discoverability* for unfamiliar participants with the multimodal interaction concept. However, it could not help users transfer their knowledge from the game to the map application due to the differences in the two contexts as mentioned by Cockburn et al. [2014]. In addition, *InteractionMap* was difficult to follow, and we guess that the location of it might be an issue as it was in the work of Srinivasan et al. [2020]. Furthermore, the observation showed a lack of enough information on how to work with gaze in general, as users struggled to find how-to confirm gaze. In addition, the co-existence of the two multimodal interactions in the same study setup caused some unexpected results for *touch-speech* in all three conditions.

4.3.10 Limitations

In this study, we faced some limitations that caused unexpected results or uncertainty in some cases. One of the most harmful limitations that we observed was technical issues. These problems were mainly with the voice system that could not listen to users or detect the correct intent. These technical issues negatively affected *learnability* of the new interactions by reducing users' trust and making them continue the tasks with the comfortable interaction techniques without willing to explore the system.

Technical issues damaged learning process for the new multimodal interactions.

In addition, some study limitations also contributed to some behaviors like rushing through the study and not exploring the new possibilities of the interface. One of the study limitations might have been tasks. The tasks were

A lack of willingness to explore new possibilities may be due to tasks and study instructions.

too easy for some participants to complete with the new interactions, or they were confusing in some cases, especially for *Jerry* and *Tom* tasks, or they did not feel natural for multimodal interaction. The other study limitation was the given instruction and the lab condition, which made the users nervous about making mistakes. Maybe using the word "successful" in the *study interface* was a bad choice that could cause this feeling.

Furthermore, there are some contradictions between the results for *gaze-speech awareness* and *touch-speech awareness* in the two different measuring methods (the awareness question and the Likert scale questionnaire). Therefore, correctly understanding the questions might have been another study limitation. For example, one participant performed *gaze-speech* and was aware of this interaction based on the questionnaire but did not name this interaction in answering the awareness question. Also, for *Game* condition, some participants answered the awareness question for both game and map application, but they answered the questionnaire only for the map application.

Chapter 5

Discussion

Based on the result of the previous chapter the two *candidate interfaces* (*InteractionMap* and *Game*), were not significantly different from *Baseline*. However, as we discussed in Section 4.3.9 "Discussion", there are some different reasons behind the good or bad results for these three interfaces. In this chapter, based on the results from the whole thesis, we discuss some possible influencing factors that might have an effect on the discoverability of the *gaze-speech* and *touch-speech* multimodal interactions.

Discoverability of a multimodal interaction might change for different tasks.

According to Oviatt et al. [1997], the likelihood of users switching naturally to multimodal interaction for a particular task depends on the type of task. However, the way we designed the tasks did not feel natural in some cases for the users. For example, we observed that the *parking* and *restaurant* tasks, when there is only one train station on the map, felt unnatural for some users. However, continuing with a multimodal interaction to, for example, navigate to one of the appeared restaurants or parking on the map came naturally to some participants. Therefore, task choice is an important factor that, in some circumstances, it might be in favor of discoverability.

In addition, the tasks that we picked could be performed by a *touch-only* interaction within two steps. 2-step tasks can be easy for a *touch-speech* interaction, as one of the participants said, "When I am touching, I already have the lo-

How a task can be performed with *touch-only* interaction would affect the *touch-speech* discoverability.

cation, or I know what I am going to do, so I don't need the speech. I continue with touch." Also, the *base system* had the MOT effect in which the context menu opens immediately by touching a POI. MOT affords for *touch-only* interaction and it can affect discoverability of *touch-speech* multimodal interaction. As such, 2-step tasks and MOT can be influencing factors on the discoverability of multimodal *touch-speech* interaction or also maybe other multimodal interactions that include touch.

Users expect an explicit confirmation of their gaze interaction.

However, the discoverability of *gaze-speech* was better than *touch-speech*. This interaction was also not easily discoverable for some participants. The biggest challenge was the lack of confirmation on selecting with the gaze. In the final study, almost all the users who performed *gaze-speech* looked at the POI (train station) during the time that they were saying the speech command. This behavior is the result of the missing confirmation for the gaze. The participants were unsure if their gaze was confirmed, so they looked at the POI until they saw the result and ensured that the system got the correct POI. Some participants tried different ways to confirm their gaze. Two participants tried pressing *space* button, some tried blinking, and some used words like "confirm" and "select." Expecting confirmation of gaze was unexpected for us. This expectation can be because of the gaze, which is a new modality, and there is no widespread convention on how to use it. On the other hand, it might be due to the similarity between this interaction and mouse interaction, especially with a *cursor* as the gaze feedback. As such, the discoverability of *gaze-speech* interaction or other multimodal interactions that include gaze might be highly affected by the discoverability of gaze, or specifically the learnability of gaze interaction.

In multimodal interactions that include speech, finding the correct speech command can be challenging.

In addition, the discoverability of a multimodal interaction that includes speech can also be affected by speech modality. In the final study, we observed that participants had problems in finding the correct speech command. Some participants did not trust the voice assistant to understand a complicated sentence that includes too much information or includes deictic references. As a result, some participants did touch to get some hints for their speech command from the menu that opens, and some participants suggested a

popup menu that tells them what to do or say next when they look at the POI. This suggestion is similar to the idea that the *interface 5* had.

In summary, besides the strengths and weaknesses that the *candidate interfaces* had for the discoverability of the two multimodal interactions, some possible external factors might have influenced the results in this thesis. Throughout the two user studies in this thesis, we could observe the effects of choice of tasks, experience with gaze modality, and confidence in speech modality as the probable influencing factors on discoverability.

Chapter 6

Summary and future work

6.1 Summary and contributions

In this thesis, our goal was to improve the discoverability of two multimodal interactions (*gaze-speech* and *touch-speech*) on a multimodal *base system*. Since we aimed to keep the habituated user experience with a touch interface unchanged, the *base system* follows the look and feel of a *touch-only* interface. Based on different approaches in the literature, we proposed seven interface designs for improving the discoverability of the two multimodal interactions in this system. Throughout two studies, we evaluated these *proposed interfaces* based on two key attributes for discoverability: awareness and learnability.

These *proposed interfaces* follow different ideas. *Interface 1* has only *POI size* gaze feedback, which is the feedback that appears when the look is on a point of interest (POI). *Interface 2* has only a *cursor* gaze feedback that shows the constant location of the gaze. *Interface 3* shows a mapping for the two possible multimodal interactions. *Interface 4* only provides information on the three possible modalities (gaze, touch, and speech) without providing any feedback on any of them. *Interface 5* uses the idea of sequential feed-

forward and tells what is possible next based on the user's first choice of modality to interact with. *Interface 6* provides a practice for the multimodal interactions in a *coloring game* context. And *interface 7* recommends the two multimodal possibilities when a unimodal interaction is detected.

Since the number of *proposed interfaces* were high, we got the help of experts who were familiar with the multimodal concepts to evaluate these interfaces. In a within-group study with seven experts, we found that *interface 3* (*InteractionMap*) and *interface 6* (*Game*) were the two best interfaces for discoverability. We picked these two interfaces as the *candidate interfaces* to proceed with. Moreover, in this study, we observed the importance of feedback for each modality in being confidently aware of them; however, it is not necessarily part of the discoverability of multimodal interactions and the scope of this thesis.

The two *candidate interfaces* received the best qualitative and quantitative results among other *proposed interfaces* in the expert study. However, they had some issues that we solved before proceeding with the final study. In *InteractionMap* we improved the design for less complexity and distraction. In *Game*, we solved the problem of knowledge transferability by providing more explanation on the purpose of the game. In addition, since the *InteractionMap* already included speech feedback on its design, we added a *speech wave* in *Game* to control speech awareness. And we combined the *POI size* and *cursor* for the gaze feedback in both *candidate interfaces*. From the experts' point of view, this new combination for gaze feedback would result in better gaze awareness for novices.

In the final study, we compared the two *candidate interfaces* with a *Baseline*. In the *Baseline*, a video shows the possible interactions and how they work. Also, the *Baseline* includes *speech wave*, and combination of *POI size* and *cursor* gaze feedback. In a between-group study, we evaluated the discoverability of *gaze-speech* and *touch-speech* in these three *interface conditions*. In the study, each user did 12 exemplary location-related tasks on the assigned condition. We measured awareness and learnability of both multimodal interactions and calculated discoverability according to them.

The results of the final study showed that the two *candidate interfaces* are not significantly different from *Baseline* in the discoverability of both multimodal interactions. In addition, *Game* performed close to *Baseline* in *gaze-speech* awareness, and better than *Baseline* in *touch-speech* awareness. Moreover, *touch-speech* awareness was significantly better in *Game* compared to *InteractionMap*. However, in general, the learnability of *touch-speech* interaction showed a bad result in all three *interface conditions*. Further analysis revealed that this bad result was mostly due to the co-existence of *gaze-speech* and *touch-speech* on an interface. When the *gaze-speech* was discovered, it was more convenient for the users, so they neglected *touch-speech*. Nevertheless, the learnability of *gaze-speech* was also a challenge in some cases. In addition to the issues that each *interface condition* had, lack of information on how to confirm their gaze was a similar issue across all of them. Furthermore, among those participants who had no prior knowledge of multimodal interaction, *Game* performed better in discoverability of *gaze-speech* than *InteractionMap*, with *Baseline* in the first place.

The key to widely adopting multimodal interactions and not keeping them as an expert-only technology is making users aware of them. Therefore, they need to be easily discoverable by novice users. We were able to identify two promising concepts for supporting the discoverability of multimodal interactions and provided valuable insights on still existing problems in the adoption of multimodal interactions. Future work can build on these results to foster the naturalness and convenience of interaction for future generations of HCI systems.

6.2 Future work

In the expert study, we excluded five interfaces. However, qualitative results, from both studies, showed that the idea behind *interface 5* could be helpful for confidently using *gaze-speech* interaction. In addition, *interface 7* was excluded because of the time limit of the thesis. In future work, these ideas are worth investigating.

In addition, in this thesis, we did the evaluation using 2-step tasks. In future work, we can assess the proposed discoverability ideas for more complex tasks that need more than two touch interactions to be finished. Specifically, this can change the results for discoverability of *touch-speech*.

Despite poor knowledge transferability in *Game*, the *coloring game* provided a good practice using a playful approach. Future works can investigate other games with either playful or gameful approaches by Deterding et al. [2011] for discoverability purposes. In addition, results showed that practice was missing in *Baseline*. Therefore, we suggest further work on different ways to embed practice into an interface.

The scope of this thesis was not finding feedback methods for *gaze-speech* and *touch-speech*. However, the observations in the two studies gave us some hints on what users expect from an interface when they perform multimodal interactions. First, users expect clear feedback on gaze confirmation, and second, they need feedback on acceptable speech commands. With the help of these observations, in the following studies, feedback on these multimodal interactions can be investigated.

Appendix A

Study Interface

Figure A.1: The *study interface* that instructed the participants through the study.

Appendix B

Expert Study

Multimodal Interaction:

Combine Look or Touch with Speech

Heuristics:

1. Awareness of 3 modalities:
   - Touch
   - Gaze
   - Speech
2. Awareness over explained multimodal interaction:
   - Combining Touch with Speech to do the task
   - Combining Gaze with Speech to do the task
3. Learnability of the (multimodal) interactions within 6 tasks.
4. Effectiveness of the information on the screen on learning the (multimodal) interactions.
5. Effect of the interface design on the general interaction experience:
   - Usability of the interface
   - Complexity of the interface
6. The metal load the interface design causes.

Figure B.1: The guideline that is used for the expert study.

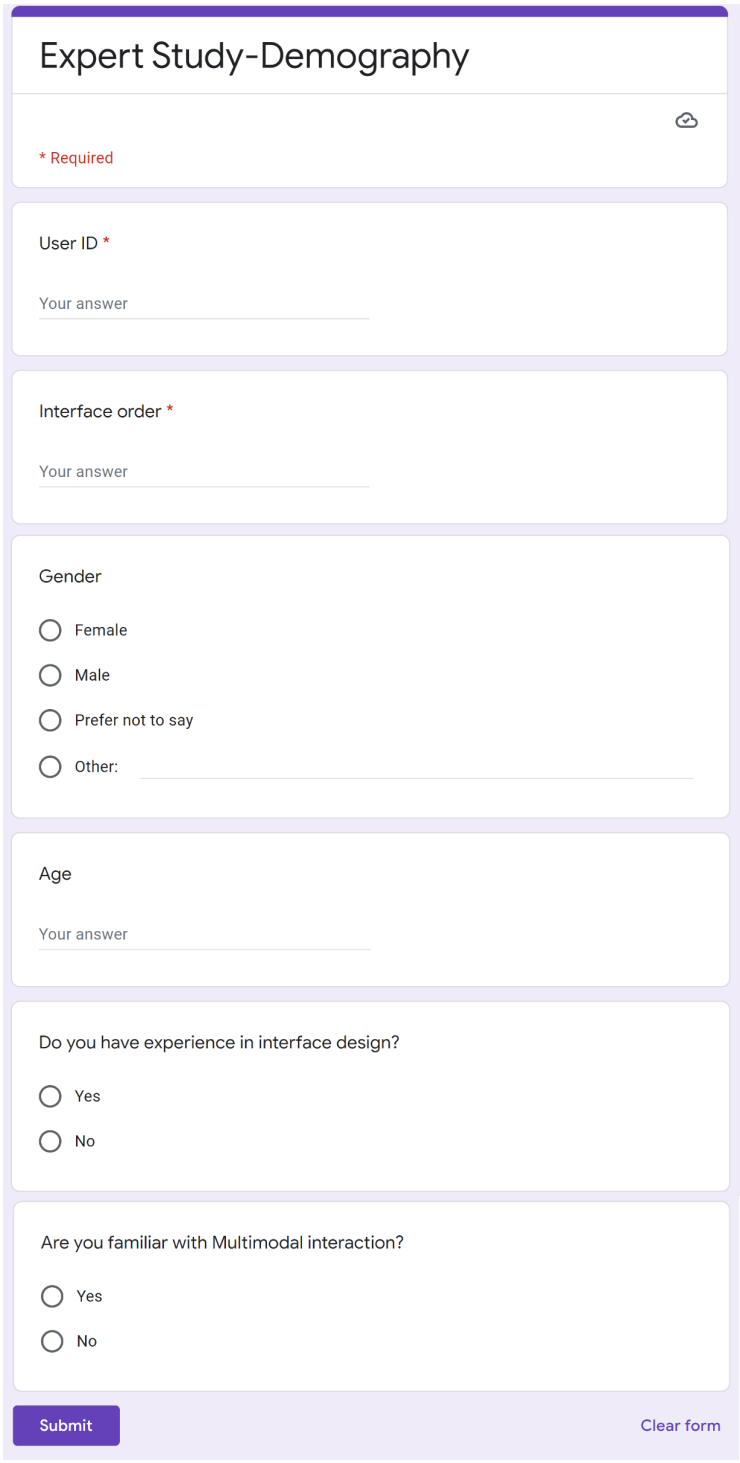

Figure B.2: Demographic questionnaire in the expert study

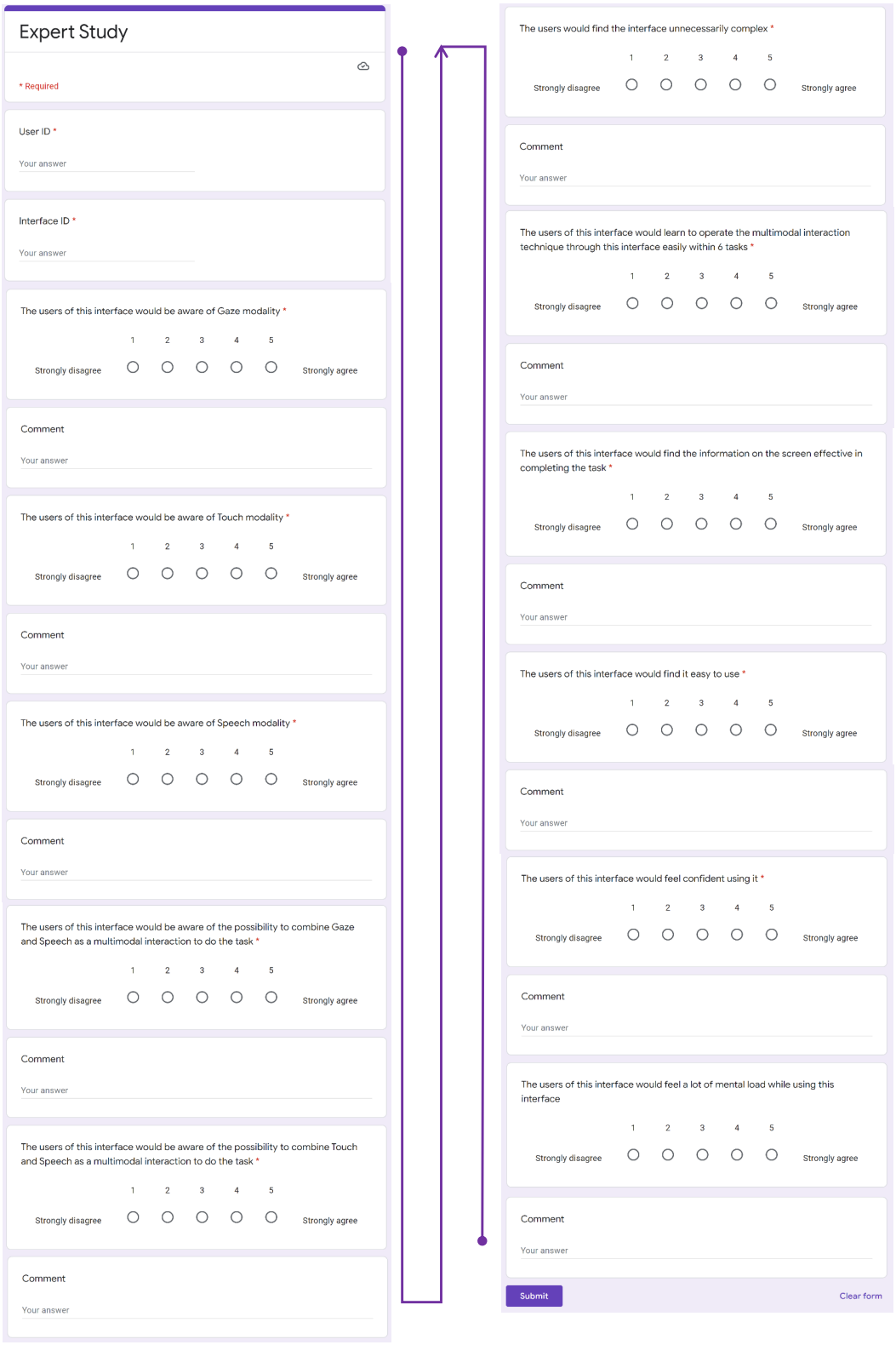

Figure B.3: Questionnaire for the quantitative evaluation in the expert study

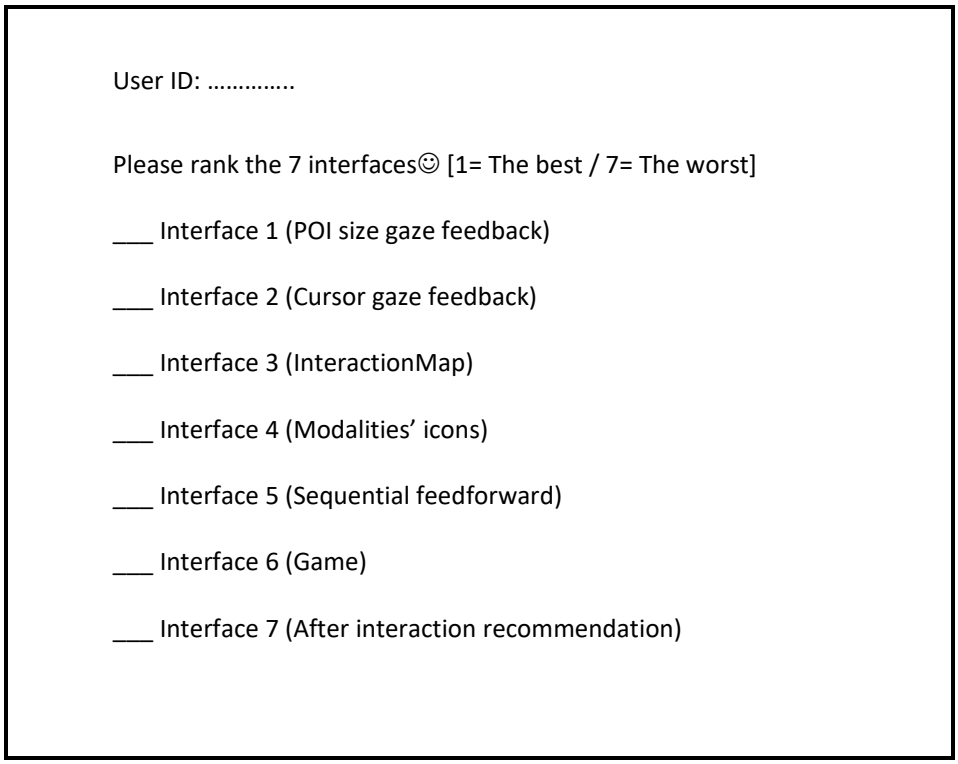

Figure B.4: Questionnaire for ranking the *proposed interfaces* in the expert study

Appendix C

Evaluation Study

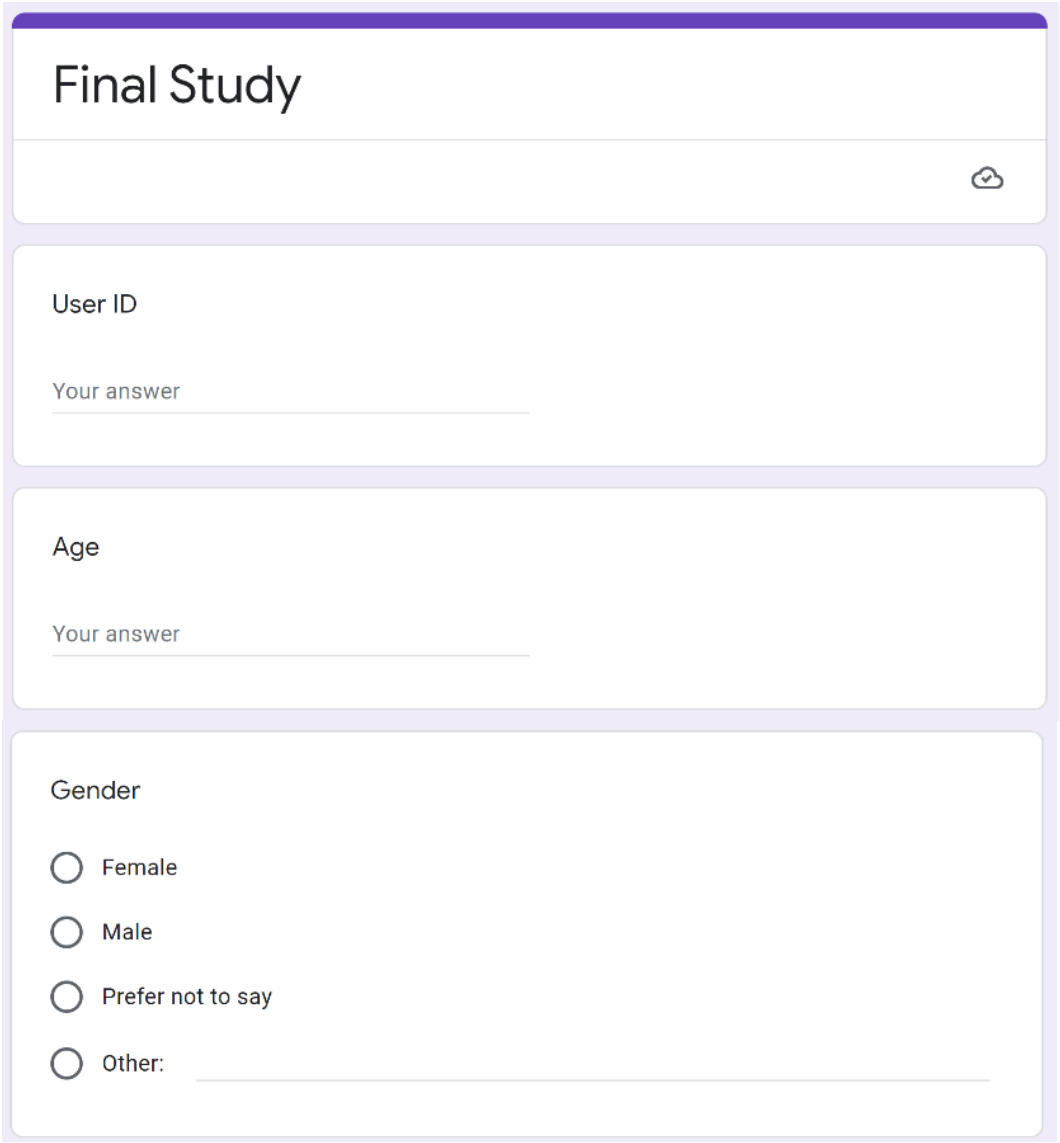

Figure C.1: Demographic questionnaire in the evaluation study

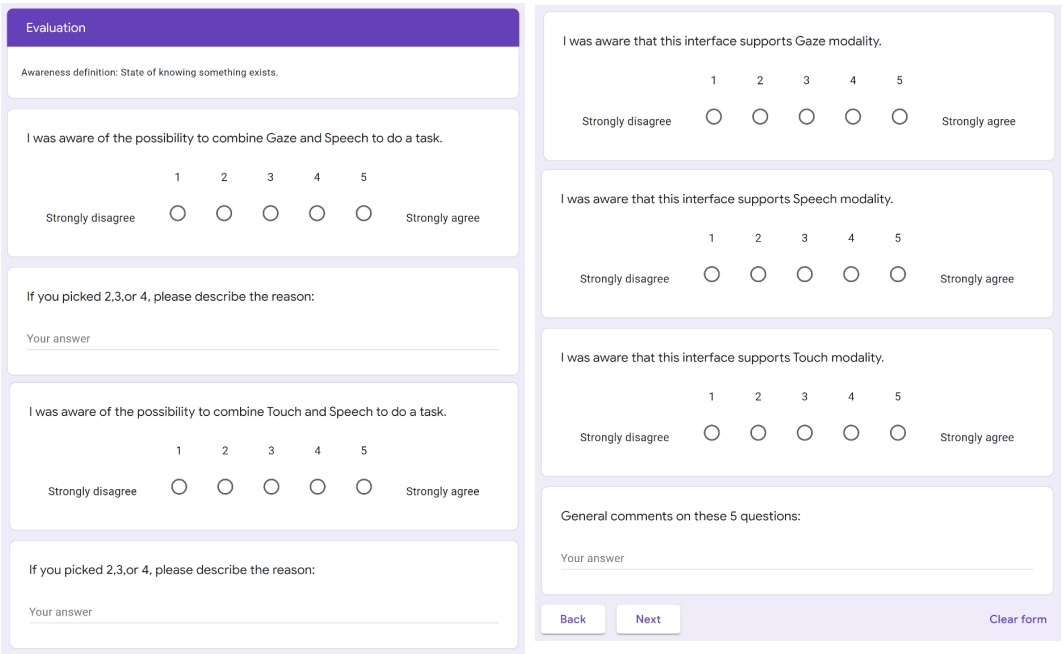

Figure C.2: Questionnaire for evaluating awareness in the evaluation study

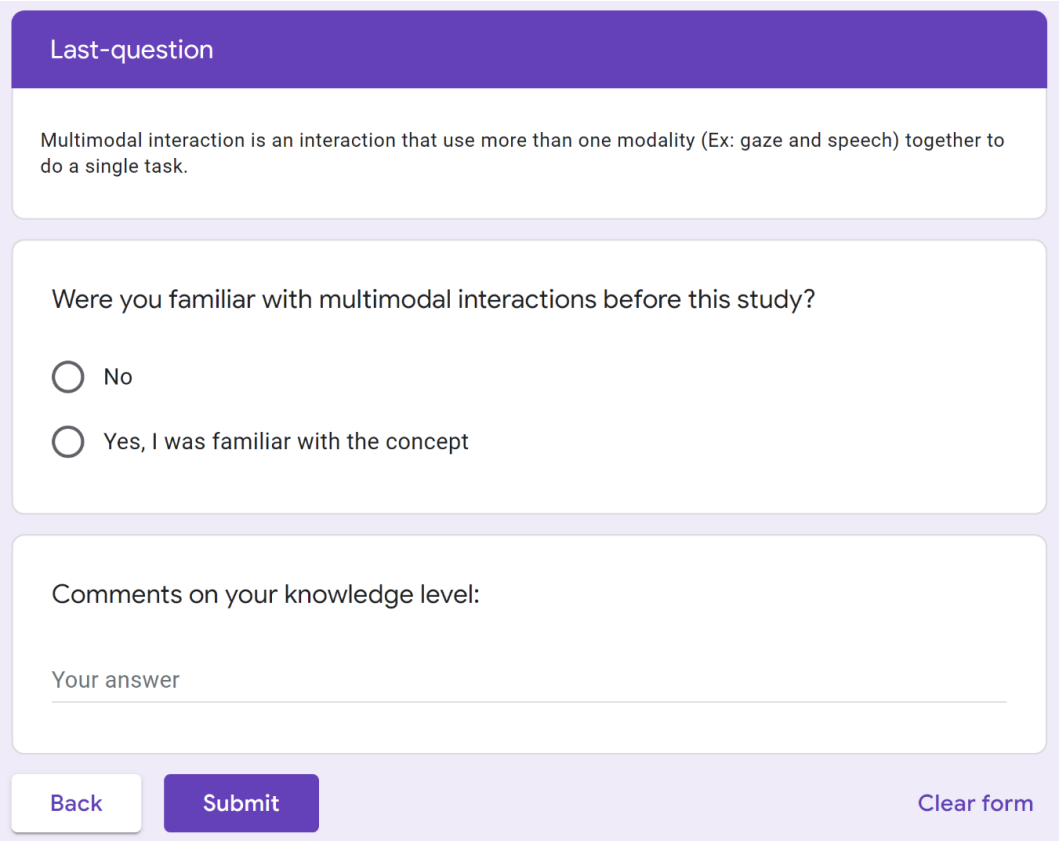

Figure C.3: Question on the familiarity with the multimodal interaction concept in the evaluation study

Bibliography

Ilhan Aslan, Michael Dietz, and Elisabeth André. Gazeover–exploring the ux of gaze-triggered affordance communication for gui elements. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 253–257, 2018.

Olivier Bau and Wendy E Mackay. Octopocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 37–46, 2008.

Jacques Chueke. *Perceptible affordances and feedforward for gestural interfaces: Assessing effectiveness of gesture acquisition with unfamiliar interactions*. PhD thesis, City, University of London, 2016.

Jacques Chueke, George Buchanan, Stephanie Wilson, and Luis Anunciação. Self-previewing gestures and the gesture-and-effect model: experimentation with responsive visual feedback for new and unlearned interactions. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) 31*, pages 1–14, 2017.

Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. Supporting novice to expert transitions in user interfaces. *ACM Computing Surveys (CSUR)*, 47(2):1–36, 2014.

Philip Cohen, David McGee, and Josh Clow. The efficiency of multimodal interaction for a map-based task. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, pages 26–27, 2000.

Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining" gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15, 2011.

Tom Djajadiningrat, Kees Overbeeke, and Stephan Wensveen. But how, donald, tell us how? on the creation of meaning in interaction design through feedforward and inherent feedback. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 285–291, 2002.

Jeff Dyck, David Pinelle, Barry AT Brown, and Carl Gutwin. Learning from games: Hci design innovations in entertainment software. In *Graphics interface*, volume 2003, pages 237–246. Citeseer, 2003.

Jinjuan Feng, Clare-Marie Karat, and Andrew Sears. How productivity improves in hands-free continuous dictation tasks: lessons learned from a longitudinal study. *Interacting with computers*, 17(3):265–289, 2005.

Paul M Fitts and Michael I Posner. Human performance. brooks. *Cole, Belmont, CA*, 5:7–16, 1967.

Dustin Freeman, Hrvoje Benko, Meredith Ringel Morris, and Daniel Wigdor. Shadowguides: visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proceedings of the ACM international conference on interactive tabletops and surfaces*, pages 165–172, 2009.

Anushay Furqan, Chelsea Myers, and Jichen Zhu. Learnability through adaptive discovery tools in voice user interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1617–1623, 2017.

Alix Goguey, Sylvain Malacria, and Carl Gutwin. Improving discoverability and expert performance in force-sensitive text selection for touch devices with mode gauges. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

Tovi Grossman, Pierre Dragicevic, and Ravin Balakrishnan. Strategies for accelerating on-line learning of hotkeys. In

Proceedings of the SIGCHI conference on Human factors in computing systems, pages 1591–1600, 2007.

Rex Hartson. Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & information technology*, 22(5):315–338, 2003.

Kay Hofmeester and Jennifer Wolfe. Self-revealing gestures: teaching new touch interactions in windows 8. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 815–828. 2012.

Slava Kalyuga, Paul Chandler, and John Sweller. Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 13(4):351–371, 1999.

David B Koons, Carlton J Sparrell, and Kristinn Rr Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. In *Readings in intelligent user interfaces*, pages 53–64. 1998.

Gordon Kurtenbach and William Buxton. The limits of expert performance using hierarchic marking menus. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 482–487, 1993.

Wei Li, Tovi Grossman, and George Fitzmaurice. Gamicad: a gamified tutorial system for first time autocad users. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 103–112, 2012.

Sylvain Malacria, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. Skillometers: Reflective widgets that motivate and help users to improve performance. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 321–330, 2013.

Darius Miniotas, Oleg Špakov, and I Scott MacKenzie. Eye gaze interaction with expanding targets. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1255–1258, 2004.

Donald A Norman. *The psychology of everyday things*. Basic books, 1988.

Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

Sharon Oviatt. Multimodal interfaces. In *The human-computer interaction handbook*, pages 439–458. CRC press, 2007.

Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 415–422, 1997.

Leah M Reeves, Jennifer Lai, James A Larson, Sharon Oviatt, TS Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, et al. Guidelines for multimodal user interface design. *Communications of the ACM*, 47(1):57–59, 2004.

Joey Scarr, Andy Cockburn, Carl Gutwin, and Philip Quinn. Dips and ceilings: understanding and supporting transitions to expertise in user interfaces. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 2741–2750, 2011.

Felix Schüssel, Frank Honold, and Michael Weber. Influencing factors on multimodal interaction during selection tasks. *Journal on Multimodal User Interfaces*, 7(4):299–310, 2013.

Nicu Sebe. Multimodal interfaces: Challenges and perspectives. *Journal of Ambient Intelligence and smart environments*, 1(1):23–30, 2009.

Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. Discovering natural language commands in multimodal interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 661–672, 2019.

Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M Drucker, and Ken Hinckley. Inchorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Matthew Turk. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195, 2014.

Jo Vermeulen, Kris Luyten, Elise van den Hoven, and Karin Coninx. Crossing the bridge over norman's gulf of execution: revealing feedforward's true identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1931–1940, 2013.

Robert Walter, Gilles Bailly, and Jörg Müller. Strikeapose: revealing mid-air gestures on public displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 841–850, 2013.

Index

Typeset December 22, 2021