HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) **Open Access**

Human-centric Computing and Information Sciences (2023) 13:18
DOI: https://doi.org/10.22967/HCIS.2023.13.018
Received: March 18, 2022; Accepted: May 19, 2022; Published: April 30, 2023

Hands-Free Presentation Tool with Co-speech Gesture Interactions: A Wizard-of-Oz Study

Ki-Young Shin[1], Jong-Hyeok Lee[2], and Kyudong Park[3,*]

Abstract

Despite active research on the design of presentation tools after the emergence of slideshow presentations, there is a lack of research findings on appropriate modalities of interactions for controlling slides in an exploratory manner. The objective of this study is to find the appropriate modality for features of controlling slides and to design usable features. This study used an iterative design process based on the Wizard-of-Oz (WoZ) prototype and participatory design (PD), which was divided into three phases. In the first phase, the participants were directly involved in the ideation process, and they created an initial design set. In the second phase, the initial prototype was evaluated by the participants with WoZ, focusing on the scope of co-speech gesture interactions. Finally, the usability of the final design set was evaluated, and it was demonstrated that the proposed design features were usable in terms of naturalness, controllability, efficiency of information delivery, and efficiency of resource use. The results also showed that verbal modality was more dominant, while many previous studies focused on creating gesture-based systems. This research is expected to provide guidance for designing a hand-free presentation with a co-speech gesture, and benefits for conducting PD research with WoZ.

Keywords

Co-speech Gesture Interaction, Participatory Design, Wizard of Oz, Slideshow Presentation Tool

1. Introduction

After the emergence of slideshow presentation programs such as Microsoft PowerPoint and Apple Keynote, the functionality and affordances of presentation slides have greatly improved. These presentation tools offer various features for computer-animated slides that allow speakers to interact with audiences and/or slides through visual effects, content, and control [1, 2]. Furthermore, there are diverse research studies examining interaction mechanisms for making presentations better for delivering speakers' messages. Some have worked on interactive navigation functions to overcome the limitation of the linear navigation structure that is predominant in the current presentation tools [3–6]. For instance, zoomable user interfaces for navigation in presentation are used in the commercial program Prezi [7–9], and other research groups have proposed hyperlinked slideshow applications [10].

***Corresponding Author:** Kyudong Park (kdpark@kw.ac.kr)
[1]Department of Convergence IT Engineering, Pohang University of Science and Technology, Pohang, South Korea
[2]Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang, South Korea
[3]School of Information Convergence, Kwangwoon University, Seoul, South Korea

The control of slide navigation is another important issue in research on presentation programs. Currently, a remote controller is the most common controller; however, it can only support a linear navigation structure. Several studies have examined alternative ways to mediate the control of slides and components. One such research used papers as the main medium for controlling slides, such as a slide card with a barcode printed on it [4], and a specially treated paper to recognize the position of a special pen [3]. Another mainstream approach is to use human gestures as a natural medium for controlling slides [11, 12]. In these gesture-based systems, certain intuitive gestures signal and perform the corresponding controls defined a priori.

Along with all these studies, diverse possible modalities are mediating the control of slides, for instance, modalities requiring handheld extra devices such as laser pointers, remote controllers, digitized papers, interactive smartboard pens, and natural human modalities such as gestures, speech, and co-speech gesture. While these modalities have their own characteristics, there is a lack of research on finding the appropriate modality for slide control. Previous research has mainly relied on the researcher's insights and heuristics for choosing the modality.

The objective of this research was to create usable, intuitive designs with features that are natural, controllable, and efficient in information delivery and resource consumption. We attempted to investigate an appropriate modality for slide control without any bias by assuming that any modality can be a candidate. Once we investigated the appropriate modality, we designed features for controlling the slides in the modality through an iterative design process. This study implemented a low-fidelity prototype and tested in Wizard-of-Oz (WoZ) setting. Creating and testing high-fidelity prototypes using the cutting-edge technologies like artificial intelligence (AI) would be time consuming and expensive. Furthermore, most derived issues during usability test may focus on the AI's performance part, making it difficult to deeply examine the experiences and needs in terms of interactions. Therefore, this research tried to move beyond technology constraints and anticipate future user reactions, allowing the design aspects to be specified in advance. These efforts can be utilized as reference data for design in the future.

2. Related Work

2.1 Studies on Diverse Modalities

Although laser pointers and remote controllers have been used for a long time to control presentation slides, many researchers have sought better modalities to enhance controllability. Interactive smartboards have been employed to improve the level of interactivity during presentations, particularly in learning environments [13]. In these systems, digital markers are used to connect users to a system.

Some researchers have used digitalized paper as a medium for presentation control [3, 4]. An early study is the Palette program, which provides control of random navigation among slides [4]. Prepared slides are printed on a special paper card with a barcode for presenters, who can change the order of the slides by reading the palette barcode using a reader. Singer and Norrie [3] suggested a different paper-based system, which allows presenters to annotate slides through special papers and a digital pen: the presenter's annotation and button-click actions are sent to the host computer controlling corresponding slides.

While these modalities are based on physical devices that allow direct controllability to users, another stream of research has focused on natural human modalities that human beings use daily to communicate, such as gestures. Human gestures have been examined as natural mechanisms for presentation control [11, 12, 14]. Cao et al. [14] evaluated three different modalities of presentation control, namely laser pointer, mouse/keyboard, and bare-hand gestural controls, using the WoZ. The research revealed that bare-hand gesture control received the highest score in all quantitative ratings on clearness, efficiency, and attractiveness. Fourney et al. [12] developed a slideshow presentation tool with gestural interfaces to resolve problems related to navigation and annotation [11] in presentation tools.

To develop these gesture-based systems, defining triggers which are the user signals of system actions that execute a functionality, is important. The triggers in many gesture-based systems are designed by system developers based on their insights and available technologies. Negroponte [15] suggested the use of human-behavior-based approaches instead of choosing easily learnable, memorable, and efficient triggers.

In contrast to the gestural modality, the verbal modality did not receive much attention for presentation systems, although it is the most dominant modality of human communication. This could be because of the considerable interruption caused by speech triggers and the reliability of speech technology. However, it might be time to consider speech as the main modality for the successful deployment of commercial speech interaction systems such as Apple Siri, Google Assistant, and Microsoft Cortana.

The combination of these human modalities forms co-speech gesture [16]. With the great interest that co-speech gesture processing has recently received in several fields including medicines [17], conversational agents [18, 19], and oral presentation [20], this modality could also be considered a candidate modality for such applications. In previous studies, however, the researchers chose one main modality, and, in that modality, they designed triggers or features for controlling slides or developed systems. This study attempted to investigate the appropriate modality for such an application in an exploratory manner, along with designing usable features.

2.2 Studies on Diverse Features of Slide Control

While previous studies targeted different modalities, they also addressed the different features of controlling slides. The Palette program targets nonlinear page navigation using special paper cards [4], and PaperPoint tackles the annotation along with nonlinear page navigation [3]. In some studies, zooming features were designed and implemented [7–9]. Diverse features have been implemented in gesture-based systems, such as linear navigation, zooming, and video control, by mapping each feature to a different gesture [11, 12, 14, 15].

3. Methodology

3.1 Research Context and Method

The main objectives of this research were to explore the appropriate modality supporting the user's natural control of presentation, which can be the basis for the design of the next generation presentation tools and control features in that modality. Because effective presentation methods may differ depending on the field and content, it is necessary to minimize the potential effect of these fields and contents. Therefore, this research was conducted with a relatively homogeneous group of engineering graduate students in a university setting.

Because of the exploratory nature of this research objective, we employed both participatory design (PD) and WoZ as the main methods for data collection. First, we employed PD in which end-users actively participate in the design process from an initial stage [21–24]. Unlike other design methods that mainly rely on researchers' insights and heuristics, PD brings together the experiences of multiple stakeholders to gather insights into different aspects of a targeted solution through their participation as co-designers. However, PD can be problematic when participants do not know exactly what they want, or how to explain their tacit knowledge. To address this issue and help the participants develop design ideas, we used a layered elaboration technique [25] that allows participants to contribute ideas provided by others while also encouraging them to expand on those earlier ideas.

Second, we used WoZ, which supports researchers in reducing the cost of full implementation through a program simulating the design features. In WoZ, users assume that they are using a real system, while a human operator reacts behind the scenes. WoZ was also used to collect users' behavioral patterns with

a target application. WoZ is widely used in study on novel gestures [26, 27] and discourse research [28–30]. Computational linguists at an early age found that the linguistic characteristics of human-machine dialogs are different from those of human-human dialogs. Because research studies on intelligent language interfaces have mostly been conducted using corpus analysis, it is critical to collect such data representing the characteristics of human-machine dialogs. Furthermore, data collected from the WoZ are useful for the development of dialog systems using data-driven machine learning algorithms because these algorithms require a realistic dataset [31, 32].

In this study, PD mainly supported initial design prototyping and design refinement, whereas WoZ was employed for usability testing (Fig. 1). First, an initial design was drawn from the participants (initial design prototyping). Many different PD techniques can be used to obtain initial design ideas. The resulting design was imitated by WoZ and tested for usability (usability test). This test, through WoZ, can give the participants opportunities to see the issues with their ideas. Once the usability test results were obtained, the design was refined (design refinement) according to usability. By dividing the entire number of iterations into three phases, we held a design workshop for each phase.

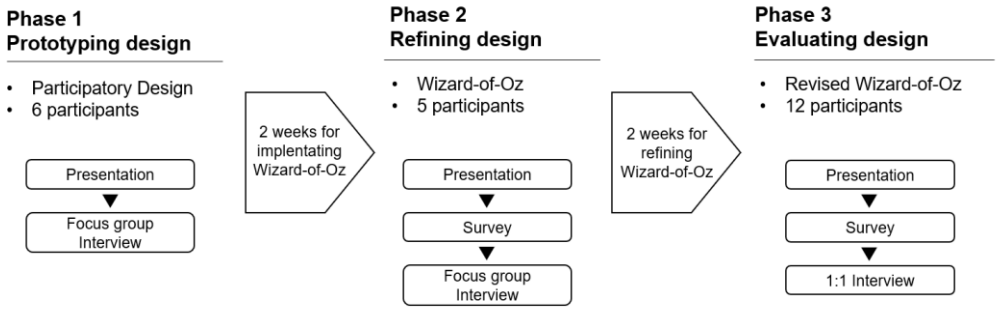


Fig. 1. Overview of the overall methodology.

3.2 Data Collection

All phases were recorded using two video cameras, and all artefacts (e.g., notes, drawings, and design results) generated by the participants were collected for analysis. In addition, the presentations made with WoZ in the second and third phases were all recorded in the raw-data format (including uncompressed IR and color images with audio from four microphone arrays) of Microsoft Kinect v2 using Kinect Studio v2.0 to develop the technology required to implement the final design features in the future.

3.3 Data Analysis

We analyzed the collected data using a constant comparative method based on grounded theory [33]. Initially, the collected data were analyzed through open coding to identify recurring themes and ideas. This representative method for qualitative data is an inductive analysis process that iteratively reorganizes categories or themes while comparing the contents of the previous data again continuously whenever a new topic appears. Recently, the method was used to analyze the behavioral and attitudinal aspects of human participants who have experienced WoZ [34], and to analyze human-agent dialogue from the point of view of recently explainable artificial intelligence (XAI) [35].

4. Phase 1: Prototyping Initial Design

In the first phase, a design workshop was held to develop an initial design prototype for the slideshow presentation tool. This phase had two primary objectives: find the appropriate modality for controlling

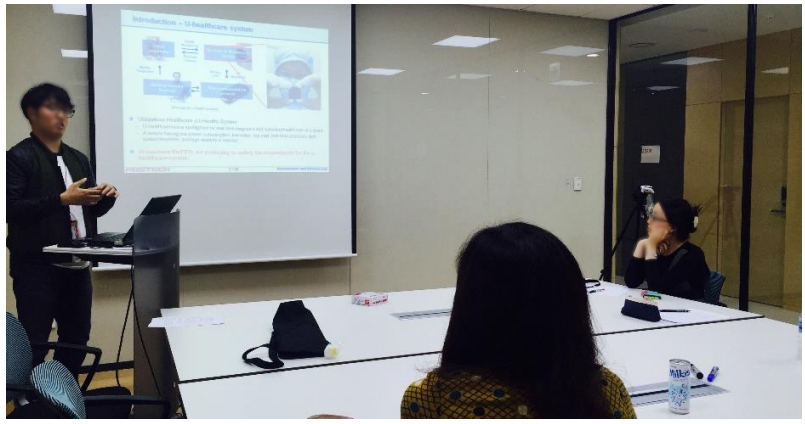
slides and derive initial ideas for designing a presentation tool. At this stage, our research approach was mainly exploratory, leaving participants in an open context where they could freely suggest and discuss ideas for the initial prototype with various features covering a wide range of presentations.

4.1 Implementation

Six participants (n=6) were recruited for the design workshop in Phase 1. The participants were either researchers or doctoral students (three males and three females) from a research university with science and engineering majors. They reported that they made an average of two work-related presentations per month. The workshop included three sessions: introduction, presentation, and focus group interviews (FGIs).

4.1.1 Introduction session

The objective of the research was described to the participants, and consent forms were collected from all participants. An ice-breaking session was then conducted to help the participants feel more comfortable interacting with each other.



(a)

(b)

Fig. 2. (a) Presentation and (b) FGI sessions of Phase 1.

4.1.2 Presentation session

The aim of this session was to help participants recall their experiences of making presentations as both presenters and audiences. Each participant was asked to make a presentation, in turn, using Microsoft PowerPoint for about five minutes, followed by extra time for questions and answers. The rest of the participants played the role of audience (Fig. 2(a)). The presenter was given a laptop and a laser pointer. The presenter could monitor the presentation material through a laptop and control the slides through the laptop's keyboard or the laser pointer’s button. In addition, the presenter could point to a

desired location on the slide using a mouse or laser pointer. The format of the presentations was mini-TED, where the participants made presentations on the topics that they had expertise in. The participants were asked to prepare their presentations in layman terms to help the audience understand the presentation. The participants were encouraged to ask questions for every presentation.

4.1.3 Focus group interview session

The FGI session was held in the form of a semi-structured interview lasting for about one and a half hours (Fig. 2(b)). First, we used a layered elaboration technique to help the participants develop design ideas. Participants were randomly divided into two groups of three participants each. Each group generated ideas for effective presentation tools based on their experiences for 40 minutes. After 40 minutes, they explained their design ideas to each other, and then the groups exchanged design ideas for improvement and elaboration. After idea generation was completed through the layered elaboration method, we conducted FGI with the participants to discuss the following two categories:

(1) Limitations of the current tools: First, the difficulties experienced by participants with current presentation tools were discussed. In this session, participants were asked to recall and associate difficulties they experienced when presenting.

(2) Suggestions for better tools: After sharing the perceived difficulties, the participants collectively generated ideas for potential solutions. The participants were told to brainstorm solutions to technical limitations and modalities, with a particular focus on making presentation interactions appear natural to both presenters and the audience.

4.2 Results

The initial design features were grouped into three categories for the main target of the interaction: designs for interaction with the audience (DIA), designs for interaction with the environment (DIE), and designs for interaction with slides (DIS), as shown in Table 1.

Table 1. Design taxonomy for initial design prototype of slideshow presentation tool

	Category	Design consideration
Designs for interaction with audience (DIA)	DIA-1. Automatic translation of the presentation for audiences with different mother tongue DIA-2. Functionality that automatically generates a QR code for an easy access to the information of the slide DIA-3. Feedback of the comprehension level of the audience by understanding the facial expressions and movements DIA-4. Listing the audiences who have questions and pressed the button	Automatic supports for bidirectional interactions between the speaker and the audience for clear delivery of information
Designs for interaction with environment (DIE)	DIE-1. Combining the slides into one for removing time waste for transitions DIE-2. Setting all equipment needed for the presentation by pressing just a single button DIE-3. Alarming the speaker through watch vibration to inform the time left DIE-4. Inclusion of videos in the file for easier management	Elimination of resource wastes for controlling environmental settings
Designs for interaction with slide (DIS)	DIS-1. Moving into the previous, next or specific page indicated by presenter's speech and gesture DIS-2. Controlling the play of videos in the slide using speech and gesture DIS-3. Showing a pointer when directed by an arm DIS-4. Zoom in/out according to presenter's speech and gesture DIS-5. Search references and contents indicated by presenter's utterance	Supports for controlling slides (medium) using natural modality both for clear information delivery, for efficient use of resources and the high controllability

First, the theme of DIA category was mainly for the clear delivery of information. The main design consideration for the features in the DIA category was to establish bidirectional common ground between the presenter and audience. Hence, the suggested features in DIA were not only for delivering information (DIA-1; DIA-2), but also for receiving feedback from the audience (DIA-3; DIA-4). The highest priority was to make the audience understand the content of the presentation. For instance, the participants wanted a system, if possible, to translate their presentation into the native language of the audience: automatic translation feature of presentations (DIA-1). Moreover, they suggested functionalities for presenters to monitor the comprehension level of the audience (DIA-3; DIA-4) to adjust the speed of presentations and the level of descriptions or to ask questions.

Second, the features in the DIE category were mainly related to the efficient use of resources. The derived ideas were mostly based on time constraints in the general context of the presentation. The design ideas in DIE-1 and DIE-2 aim to minimize the time transition between presentations and equipment settings. When there are multiple speakers, complying with the time limit is important, but it is challenging to monitor the amount of time left. DIE-3 represents this situation, allowing the speaker to know the time left for his presentation by alarming. Some participants experienced troubles with video play when the video file was not copied correctly or the file directory structure did not match. They suggested that videos in a cloud service could reduce the difficulty in managing video files in presentations (DIE-4).

Finally, the ideas in the DIS category were both for the clear delivery of information and the efficient use of resources. The slide view, including slide pages, animations, and video content, should be changed with appropriate timing. DIS-1 and DIS-2 address this need because their fundamental objective is to reduce the time taken to search a certain slide. This is not only to control the use of time but also to deliver information more clearly by minimizing interruptions from delays. The participants also suggested certain features to highlight and emphasize the main points during the presentation. DIS-3 allows the presenter to directly indicate the main point by showing a pointer and DIS-4 emphasizes controlling the spatial use of the screen. DIS-4 was used to control the slide view for the efficiency of the spatial use of the screen. Interestingly, all the features in the DIS category were related to user interfaces through the presenter's speech and gestures, which are communicative modalities for expressing knowledge, thoughts, and intentions. Hence, the participants perceived that speech and gestures would be the most natural way of expressing their intention to control the presentation slides.

Among the three categories derived in the initial design prototype of a slideshow presentation tool, we narrowed down the design scope into the DIS category, as our focus was to design features for controlling slides, and the main modality we targeted was co-speech gesture. Although we identified our target modality, we conducted a deeper analysis of the nature of the modality for the features developed in Phase 2.

Therefore, the key features in DIS refer to 'co-speech gesture interfaces for controlling slides.' With this focus, we extended the set of features by adding two more features: to use co-speech gesture to enter and exit hyperlinks, and to change the properties of a selected object. DIS-5 was removed for the final set of features because it would be more appropriate in a web search interface than the presentation itself.

Table 2 presents the final set of features for Phase 1. For most features, their triggering speeches or gestures were not defined in the first phase because the participants focused on a high-level design taxonomy. The tentative triggers were created. We used a predefined command-like speech or gesture which we call 'guided gestures or speeches' to make the triggers, and all the speech triggers of features were in Korean.

5. Phase 2: Refining Design

The main aim of the second phase was to help participants concretely experience and refine the usability of the key features derived in Phase 1. Central to this refinement process is the realization of the initial design through WoZ.

Table 2. Final list of the co-speech gesture feature set obtained from Phase 1

Features	Triggers
Page navigation (DIS-1)	Speech Triggers: *'next', 'previous',* *'move to <SLIDE_SUBJECT>',* *'go to <RELATIVE_PAGE_INDEX>',* *'go to page <PAGE_NUM>'* Gesture Triggers: *swiping a hand*
Video control (DIS-2)	Speech Triggers: 'play', 'stop', 'pause', 'forward', 'backward', 'next bookmark', 'previous bookmark' Gesture Triggers: *pointing for specifying a video*
Pointer (DIS-3)	Gesture Triggers: *pointing to a spot on the screen*
Zooming (DIS-4)	Speech Triggers: *'zoom in', 'zoom out',* *'zoom in here'* Gesture Triggers: *pointing for specifying a spot to zoom in*
Hyperlink (added)	Speech Triggers: *'enter the hyperlink', 'exit',* *'enter this hyperlink'* Gesture Triggers: *pointing for specifying a hyperlink* *scrolling for navigation in the link page*
Object control (added)	Speech Triggers: *'enlarge this picture',* *'reduce the font size',* *'move this image there',* Gesture Triggers: *pointing for specifying an object* *pointing for moving the selected object*

5.1 Implementation

It was not an easy task to realize all the features suggested by the participants in Phase 1 because of the necessity to employ high-end artificial intelligence technologies, such as automatic speech recognition, (multimodal co-gesture) spoken language understanding, and dialog management (system action planning), for processing the presenter's co-speech gestures. Although these are emerging research fields [36], some of them are yet to be applied to a complete system. It is time-consuming, data-intensive, and expensive to implement all the features for testing.

Instead, we built a WoZ by implementing an extended set of co-speech gesture features. Five participants (n=5) from the first phase continued to participate in this phase; one participant dropped out for personal reasons. Trial, presentation, survey, and FGI sessions were conducted during this phase. In the trial session, the participants simply tried out the target features through WoZ because they were not familiar with the new interface; then, the participants made their presentations with WoZ only using the co-speech gesture features (Table 2). Unlike Phase 1, no laser pointer was given. All these presentations were recorded using Microsoft Kinect v2 for data collection, which could inform the later development of the technical features such as skeleton tracking [37] and motion captures [38]. By demonstrating some technical functions, we ensured that the participants believed that the features were implemented, so the collected data reflected the nature of human usage of the features.

It is possible that the proposed features seemed useful during the design process but were not perceived to be useful when tested in practice. Because WoZ allows the participants to experience the usefulness of the design features, they discussed whether a certain feature should remain in the list, be eliminated, or be modified based on the trials of the features. To examine the usability of the features, we conducted a survey. The most widely used to evaluate usability is Nielsen's 5 quality components (learnability, efficiency, memorability, errors, and satisfaction) [39]. In this study, errors were not considered because they did not occur at all due to WoZ settings. At this stage, overall satisfaction was investigated in depth in the interview session rather than survey because it is formative evaluation rather than summative evaluation. Thus, we developed four main aspects of usability: how natural it is to interact using the feature (naturalness), how easy it is to control slides (controllability), how much it increases the efficiency of delivering messages (efficiency of information delivery), and using resources such as space, time, and human efforts (efficiency of resource use) on a five-point Likert scale (1–5; higher score is better). The participants were asked to respond to questionnaires individually. Then, an FGI session was held to discuss the following issues regarding the target features.

5.1.1 Complete descriptions of features

Since the design ideas derived in Phase 1 were still abstract, the participants were asked to create detailed descriptions of the interactions, particularly answering the following questions:

Should the speech and gesture be natural or guided?

Should the interactions be unimodal, multimodal, or both, depending on the context?

What would be the detailed triggering of speech or gestures if they should be guided?

5.1.2 Additional design features

The participants were asked to supplement the feature set with additional features inspired by the trials in this phase.

5.2 Data Analysis

In this phase, we obtained the complete specifications of the design features (Table 3) with usability scores (Table 4).

Table 3. List of final design feature specification

	Verbal modality	Gestural modality	Simultaneous multimodality	Natural/Guided
Page navigation	Primary	Primary	Yes	Both
Video control	Primary	Auxiliary	Yes	Both
Pointer	Not used	Primary	No	Natural
Zooming	Primary	Primary	Yes	Both
Hyperlink	Primary	Auxiliary	Yes	Both
Object control	Primary	Auxiliary	Yes	Both
Annotation	Primary	Primary	Yes	Both

First, the modalities of the design feature were coded as "primary," "auxiliary," or "not used." "Primary" is used when the corresponding modality is to express the main intention for a feature, and "auxiliary" is used when the modality complements other modalities for communicative purposes. For example, a speaker may say "go to this link" with a gesture pointing to a hyperlink expressing his intention to open the hyperlink. Here, the verbal modality is mainly used to express the speaker's intention to open the hyperlink, whereas the gestural modality complements this intention in a spatial dimension, indicating the position of the target hyperlink, which is not inferred from the utterance of the verbal modality. For the video control feature, the verbal modality is "primary," and the gestural modality is "auxiliary." When

both verbal and gestural modalities are coded as "primary," it means that there are two cases where in one modality is used primarily while the other is used auxiliary, and the other way around.

Second, we coded the features according to their simultaneous use of two modalities, that is, whether the feature requires two modalities simultaneously. Finally, the triggers were coded into two different types: guided and natural. A guided trigger is a command-like trigger created by developers. Commands like "next," "play" and pre-designed gestures mapped to a system action are guided triggers. On the other hand, natural triggers are in the form of natural utterances or gestures used in daily life. An utterance "let's look at the detailed procedure on the next page" can be an example of a natural trigger in the page navigation feature.

Table 4. Result of usability scores in Phase 2 (P2) and Phase 3 (P3)

	Naturalness		Controllability		Efficiency on information delivery		Efficiency on resources use	
	P2	P3	P2	P3	P2	P3	P2	P3
Page navigation	4.1 (1.0)	4.3 (0.9)	4.2 (0.5)	4.3 (0.9)	3.4 (1.1)	4.6 (0.7)	3.7 (1.3)	4.4 (0.7)
Video control	3.0 (1.3)	4.4 (0.7)	3.2 (1.1)	4.3 (0.7)	4.2 (0.7)	4.2 (0.9)	4.6 (0.7)	4.6 (0.6)
Pointer	4.4 (0.7)	4.3 (0.6)	3.8 (1.3)	4.3 (0.7)	4.4 (0.4)	4.4 (0.7)	4.4 (1.1)	4.3 (0.7)
Zooming	3.6 (0.4)	4.3 (0.6)	4.0 (0.8)	4.4 (0.8)	3.8 (1.1)	4.6 (0.8)	3.2 (1.3)	4.6 (0.8)
Hyperlink	3.4 (0.9)	4.7 (0.5)	3.4 (1.1)	4.3 (1.0)	3.4 (1.4)	4.1 (1.1)	4.0 (0.8)	4.1 (1.1)
Object control	2.0 (0.8)	3.7 (0.8)	2.4 (0.7)	3.9 (0.6)	2.4 (1.4)	3.6 (0.8)	3.8 (1.3)	3.6 (0.9)
Annotation	-	3.5 (0.8)	-	3.9 (0.9)	-	3.6 (0.9)	-	3.8 (0.9)

Values are presented as mean (standard deviation).

Here, we briefly describe the specifications of each feature in the final set tested in Phase 2.

The page navigation feature consists of two functionalities: linear and random. For linear navigation, guided swiping hand gestures and guided speech were used. Natural verbal triggers were added to the final set in the second phase. For random navigation, only guided speech was originally available. Natural speech is included in the second phase. In natural speech, pages can be specified by their properties or content.

The video control feature includes three functionalities: playing, pausing, and moving to bookmarks. In the initial WoZ system, only guided verbal triggers were included with pointing gestures to optionally specify a target video. In the second phase, participants added natural triggers to these functionalities. "Naming bookmarks" was also added to support natural triggers by calling the bookmarks by their appropriate names depending on the contents instead of using the word "bookmark," which interrupts a presentation.

The pointer feature is formed only by pointing gestures. When the presenter points to a spot on the screen, a pointing gesture is enabled.

The zooming feature includes zooming in and out of functionality. Two different types of guided verbal triggers were covered. The pointing gesture was used together with a verbal trigger when the zooming function was used in a specific spot. In the second phase, natural triggers were added for the other features and stretch and pinch gestures without any verbal command were added for zoom in and out functionalities, respectively.

The hyperlink feature includes the opening and closing of a hyperlink. The triggers for this feature were guided verbal triggers with pointing gestures in the initial design. Natural verbal triggers with pointing gestures are then added.

The object control feature includes functionalities such as moving objects and changing the size of the objects on the screen. Natural verbal triggers with the pointing gesture were added to the initial design, which included only guided verbal triggers with the pointing gesture.

The annotation feature was newly added in the second phase and allowed the user to annotate the slide to emphasize their points. This feature comprises two components: mode transition and annotation. For

presenters to use the annotation feature, they must switch from the normal mode to the annotation mode through guided and natural verbal triggers. Once the mode is set to annotation mode, the presenters can use natural pointing gestures for annotation.

5.3 Results

From the FGI session of Phase 2, the participants primarily suggested speech-based, gesture-supported, natural, and simultaneous multimodal interactions.

5.3.1 Speeches vs. gestures

The participants in this study agreed that gestures alone were not enough to form a natural and usable interaction from the first phase. Gestural expressions are better suited for expressing shapes and directions, but it is challenging to express other types of messages, such as “play the video.” By contrast, verbal expressions can easily contain these messages. One participant said, “It would be very hard to express yellow through gestures.” Furthermore, what the participants suggested was not that the verbal modality supports the gestural modality but the other way around, that is, the verbal modality dominates. This is similar to that discussed in [16]; they showed that gestural modality mostly supplements verbal modality in spatial dimensions. In the design specification (Table 3), four of the features (page navigation, pointer, zooming, and annotation features) used the gestural modality as the primary, while the others mainly used the verbal modality, obtaining the position of the target object from the gestural modality.

Both primary and auxiliary gestures were predominantly in the form of pointing gestures. The difference between the primary and auxiliary use of the gestural modality was thus based on whether the pointing gesture alone could express the presenter’s intention. The features tagged with the primary use of pointing gestures were the pointer and annotation features. If the speaker points at the screen, he already expressed his intention to emphasize a point or write on the screen. However, in the case of the annotation feature, the pointing gesture is not the only primary trigger. To enable the pointing gesture as the primary trigger, the speaker needs to verbally express that he/she wants to use the annotation feature (e.g., “use annotation feature”), allowing the system to distinguish what the intention of the pointing gesture would be. For this purpose, the annotation feature requires verbal and gestural modalities as the primary. The video control, zooming, hyperlink, and object control features, on the other hand, use the pointing gesture to complement the intention by specifying the target object leading the corresponding tag to be “auxiliary.”

There are two exceptional features that use gestures other than pointing gestures: page navigation and zooming. Page navigation does not require any complementary information from pointing gestures; therefore, verbal modality alone becomes the primary modality. However, a swiping gesture is suggested as a guided trigger; therefore, the gestural modality can also be used as the primary modality. Here, two modalities are exclusively used to express the intention of page navigation; that is, either modality can be used alone. With the zooming feature, the pointing gesture has already been used as an auxiliary modality to specify the target position of zooming. In addition, the participants used a guided trigger for the zooming feature using hand pinch and stretch gestures. Because the gestures for this trigger can only convey the intention of the speaker to zoom in/out, the gestural modality of the zooming feature is regarded as primary.

5.3.2 Natural vs. guided

In gesture-based systems proposed in previous studies [11, 12], presenters are required to perform specific guided gestures for triggering system actions mostly designed by system developers. This study explored what type of gestures and speech would be the most appropriate triggers for such applications, instead of designing guided gestures a priori. Our participants’ ideas about triggers were simple: “as natural as possible.” Presentations are real-time performances, where the main objective of the presenter is to deliver a message to the audience. If the presenter makes an unnecessary gesture or speech in the

middle of the presentation, the interruption disturbs the message delivery. Consequently, our participants often expressed their intention to trigger a system action naturally during a presentation in advance; therefore, the most preferred interface would come from understanding the intentions from these natural expressions. In terms of speech interaction, they wanted the system to understand natural phrases such as "Well, then let's watch the video, here" instead of constraining presenters to use only guided triggers such as saying "play."

However, participants did not want to eliminate all command-like guided triggers. One of the participants, in the FGI session, said 'However, it would be also strange to keep saying "let's move to the next page' again and again." Although there are many possible sentences with the same intention, repeatedly saying naturally formed sentences can cause even a higher level of interruptions than those caused by the continuous use of guided speeches or gestures. Therefore, the mixed-use of both natural and guided co-speech gesture was suggested to be the most suitable for such presentation applications.

There are several implicit factors to consider regarding natural verbal triggers. Natural speech utterances include sentences with multiple intentions. One participant brought this issue into the discussion. When presenters naturally express their intentions, they may use complex sentences to express multiple intentions. She explained a situation, "let's watch the video again after moving to the page with the video." This requires more challenging natural language processing technologies. The dependent clause 'after moving to the page with the previous video' shows the intention, which is to perform page navigation. However, the target is specified by its properties, which are highly dependent on the context.

Overall, this aspect is closely related to the domination of verbal modality over gestural modality in presentation control. Since the most natural way of expressing they intend to use verbal modality, verbal modality dominates in the design feature specification, and gestural modality is mainly auxiliary for supplementing the spatial dimension of the message.

5.3.3 Usability of intermediate design features

Although the number of samples was small (n=5), we examined how the participants perceived the usability of the features (Table 4). First, the controllability of all features, except object control, received high scores. This means that these design features allow presenters to easily control the slide, as desired. However, two of the participants commented that the delay in the system reaction that occurred in the video control made them feel uncontrollable. The other two participants had low scores (2 points) for the controllability of the hyperlink feature owing to the absence of control after entering a hyperlinked page (For this, we added a scrolling function to a website. Other than that, we thought adding more features for controlling websites would be out of scope). These two comments show the importance of the response time of the system and coverage of the features.

Second, we could find how natural it is to use gestures for emphasizing (naturalness, 4.4), and the presenter points to deliver information more clearly (efficiency of information delivery, 4.4). Notably, the efficiencies of information delivery and resource use for video control features are high (4.2 and 4.6%, respectively). All the participants mentioned that they would use the video control feature even if only guided gestures were supported because there was no available alternative for controlling videos from a distance. The hyperlink feature, which does not have any alternatives, also received slightly lower scores because two participants focused on the absence of further control on the linked page.

Finally, it is also remarkable that while the first three scores for the object control feature were quite low, the participants suggested leaving it in the final design list. Since presentations usually made by the participants are often academic and instructive with adult audiences, they would hardly use the object control feature, yet there can be other types of presentations, such as those used for teaching children, in which this feature could be helpful; therefore, the feature was retained in the final list for Phase 3.

6. Phase 3: Evaluating Design

The final design features were tested in the third phase. By implementing the final design as a WoZ system, we could see how usable the final product would be and how feasible it would be to implement the final design with the currently available technology, even before the development of full-scale actual technology.

6.1 Implementation

We implemented the WoZ of the final design product and tested the usability of the features in terms of their naturalness, controllability, efficiency of information delivery, and efficiency of resource use, as in Phase 2. We also collected data on the recorded presentation for system development because they could reflect the user behaviors expected in the actual implemented system.

All five participants from Phase 2 joined the final test, and we collected seven additional subjects (six males and one female, all graduate students) to enlarge the sample size (n=12). Each subject made a five-minute-long presentation with WoZ setting as in Phase 2 and listened to the presentations made by the other subjects. After the presentations, the subjects completed a survey measuring the naturalness, controllability, efficiency of information delivery, and efficiency of resource use. The topic for the presentations in this phase was not pre-assigned, as long as they shared their knowledge with the audience, as the presentations in the first and second phases. After the usability test, we conducted semi-structured 1:1 interview with the participants who attended all three workshops to complement the findings from the quantitative survey and collect feedback on the overall design process.

6.2 Results

The results of the usability tests are listed in Table 4. Overall, we observed an improvement in all four categories based on the usability scores of the intermediate design features. Specifically, the naturalness and controllability of the features improved significantly during this phase. This improvement comes from allowing natural triggers. According to our participants, the presenters commonly express their intention to play a video, zoom in, or open the linked site because it is a sudden event that the audience cannot expect in advance. Therefore, the naturalness and controllability of these features were improved. However, a less sudden event such as page navigation did not improve much in terms of naturalness and controllability. The presenters rarely expressed their intention to change pages forward or backward using a natural trigger and, in most cases, used guided gestures instead. The participants mentioned that natural speeches for changing slides were used only to fill the gap during the transition of pages. In other cases, saying something only for transition disturbs, hence leading the presenters to use gestures instead. The participants also emphasized the advantage of using gestures for page navigation. In other words, the presenters can explain while navigating the pages. However, natural triggers seem effective for the page navigation feature when the presenters randomly access pages by changing the screen among pages far from each other. Random accesses, such as re-visiting the page long before, were sudden effects, making the presenters naturally express their intention. Because of the addition of natural triggers, the presenters did not need to express any extra guided triggers, causing fewer interruptions and increasing the efficiency of information delivery and resource use.

It is also notable that the scores on the efficiency of information delivery and resource use for the zooming feature were largely improved, partly because of the additional guided triggers added in the second design workshop. As described earlier, more guided triggers, apart from natural triggers, were added for the zooming feature: pinch and stretch gestures. Almost half of the cases using the zooming feature in their presentations used these guided gestural triggers instead of natural speech-driven triggers. One participant commented, “The pointing gesture and zooming gesture were not that unnatural because they are quite intuitive gestures.” It appears that the common past experiences and intuitiveness of such gestures allow the speaker and audience to accept guided gestures. Moreover, the participants often felt

more comfortable using carefully designed guided gestures than natural or guided speech because they could interact with the slide while verbally delivering their messages.

Finally, the results showed considerably lower scores for the object control and annotation features than the scores of the others in all four measures. We speculate that these two features were uncritical functionalities for the type of presentation that the participants made in this research context (e.g., academic and instructive presentations for adult audiences). None of the presenters used the object control feature, whereas the others used the annotation features. Because these features dynamically change the contents of the slide, the need for these features comes from situations where it is critical to have active interactions between the content and the audience.

7. Discussion

This section further discusses some implications of the study, such as the use of the designed features with other existing tools, and the benefits of using WoZ in the design process.

7.1 Mixed Use with Other Tools

An important observation from individual interviews is that the proposed design is not always the optimal solution for interacting with slides during a presentation. A remote controller is suitable for linear page navigation because it can provide the presenter with the ability to navigate seamlessly and linearly to the audience. Regarding the guided gestural trigger for linear page navigation, the remote controller also allows the presenter to change the slide while explaining with less interruption owing to the absence of visible gestures. A laser pointer can provide the same effect in a simpler manner.

However, the biggest limitation is the need for extra devices; to use these functionalities, presenters should always bring these tools with them. When the presenter does not have tools with them, the proposed features can be used as alternatives. The linearity of the remote controller is another limitation that can be determined by the page navigation feature. Finally, the pointer feature reinforces the visibility, which is a limitation of the laser pointer. The laser pointer was not visible when the distance between the screen and audience was large.

Because the participants prioritized making the least interruption for a better presentation, they suggested the mixed-use of the proposed design features with the currently available tools because they can complement each other. While a remote controller provides seamless linear navigation, the proposed design features provide the ability to access random pages, control videos, enter hyperlinks, etc. The choices among the tools and features were made in real time based on the expected interruption level of the possible features.

7.2 Benefits of WoZ

During the individual interviews, the participants agreed with the helpfulness of WoZ in the design process. In this study, the utilisation of WoZ in the PD framework resulted in two main benefits: design process and data collection. The former implies that WoZ helps participants concretise their ideas better, and the latter implies that WoZ can reduce the cost of data collection by integrating the data collection process into the design process.

7.2.1 Benefits to design process

In this study, the integration of WoZ and PD makes the refinement process meaningful. The target domain of our study, co-speech gesture interaction, is an exemplar case in which the implementation of the system is highly laborious, costly, time-consuming, and technically challenging. Hence, it is not feasible to implement all the design features suggested by participants in the PD process. Instead, the

WoZ can function as a simulation of a real system. In this study, we observed that the utilization of WoZ was helpful in generating design ideas and deepening the details of each feature in the design process.

7.2.2 Benefits to data collection

Apart from the benefits to the design process, WoZ also provides benefits in terms of data collection. In this study, the implementation of co-speech gesture interactions requires several challenging technologies, such as co-speech gesture recognition and dialog management for planning system actions. However, collecting a realistic dataset for natural co-speech gesture is complex. This collection process is more than simply collecting natural co-speech gesture in a presentation setting because it is known that human beings use different languages according to whether they interact with machines or humans [28]. To manage this difference, WoZ has been utilised to collect realistic natural utterances for human-machine interactions in language processing fields [28, 30–32, 40]. If one has a design of intelligent interactions and wants to implement it, then one probably needs a WoZ system to collect a realistic dataset for training a machine-learning model.

In this study, we integrated WoZ into the design process and used it as part of the design process. As a result, we did not need a separate data-collection process. Instead, we collected data during the design process by recording the participants' trials using Microsoft Kinect v2. The resulting dataset contains a total of 17 video recordings comprising about two and a half hours occupying 1.5 TB. The dataset was multimodal, including 1920×1080 uncompressed color videos (30 Hz), 512×424 uncompressed infrared videos (30 Hz), and audio streams from a microphone array with four microphones.

8. Conclusion and Future Directions

In this study, we conducted an iterative PD process with WoZ to design a new slideshow presentation tool with co-gesture-speech interactions. The overall process was divided into three phases. In the first phase, the participants brainstormed new features for a slideshow presentation tool and created an initial design set. Among the diverse features in the initial design set, we confirmed that speech and gestures would be the most suitable form of interaction between the presenter and slide. Focusing on the scope of co-speech gesture interactions, the initial design was then evaluated by the participants in a WoZ setting, allowing us to have a refined final design in the second phase. The refined final design set reflects the need for both natural and guided triggers for each feature. We also observed that verbal modality was more dominant, while many previous studies focused on creating gesture-based systems. Finally, the usability of the final design set was evaluated, and it was demonstrated that the proposed design features were usable in terms of naturalness, controllability, efficiency of information delivery, and efficiency of resource consumption.

In this study, we obtained a set of design features for a new slideshow presentation tool using co-speech gesture and could unpack the nature of presenters' co-gestural speech as an interaction medium. Nevertheless, this study has several limitations. Statistical generalization is difficult because the number of subjects was small, and it consisted only of engineering-related majors. In addition, the findings of this study are limited to the slideshow type presentation. In this study, only some elements of usability were evaluated, but if the types of gestures increase, evaluation of memorability and learnability aspects should also be treated as important. It will also be necessary to evaluate interactions at the user experience level including affection.

We present the possible directions for future research. Because our application targets slideshow presentation support, we restricted the scope of the functionalities to on-slide functionalities. The only exception was the scrolling functionality on a website opened through a hyperlink on a slide. The participants suggested some features beyond this restriction, such as complex webpage control and annotation functionalities. In future research, to extend the application to an interactive smartboard through co-speech gestures, one might need in-depth research on the nature of co-speech gesture usage for these features.

We plan to work toward the develop the technologies required for the implementation of the design. There are still immature technologies required for the implementation of the design, such as multimodal speech and gesture understanding and target-pointed on-screen object calibration. Using the dataset collected through WoZ in this study, we plan to improve these technologies and develop next-generation presentation tools for seamless and natural interaction. In particular, with the advent of tools such as Google ML Kit and Huawei ML Kit, the gesture recognition and automatic speech recognition algorithm has become lighter, enabling high-accuracy and high-speed recognition even with mobile phones. By taking full advantage of these advantages, we plan to develop a tool that can perform hand-free presentation using only with a mobile phone's camera sensor and microphone.

Author's Contributions

Conceptualization, KYS. Funding acquisition, KP. Investigation and methodology, KYS. Project administration, KYS. Resources, KYS, JHL. Supervision, KP. Writing of the original draft, KYS. Writing of the review and editing, KYS, KP. Software, KYS, JHL. Validation, JHL, KP. Formal analysis, KYS, KP. Data curation, KYS. Visualization, KP.

Funding

This study was supported by the Translational R&D Program on Smart Rehabilitation Exercises (TRSRE-Eq01), National Rehabilitation Center, Ministry of Health and Welfare, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1G1A1012063). The present research has been conducted by the Research Grant of Kwangwoon University in 2020.

Competing Interests

The authors declare that they have no competing interests.

References

[1] D. E. Zongker and D. H. Salesin, "On creating animated presentations," in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, San Diego, CA, 2003, pp. 298-308.

[2] L. P. Rieber, "Animation in computer-based instruction," *Educational Technology Research and Development*, vol. 38, no. 1, pp. 77-86, 1990.

[3] B. Signer and M. C. Norrie, "PaperPoint: a paper-based presentation and interactive paper prototyping tool," in *Proceedings of the 1st International Conference on Tangible and Embedded Interaction*, Baton Rouge, LA, 2007, pp. 57-64.

[4] L. Nelson, S. Ichimura, E. R. Pedersen, and L. Adams, "Palette: a paper interface for giving presentations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA, 1999, pp. 354-361.

[5] R. Spicer, Y. R. Lin, A. Kelliher, and H. Sundaram, "NextSlidePlease: authoring and delivering agile multimedia presentations," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 4, article no. 53, 2012. https://doi.org/10.1145/2379790.2379795

[6] T. Moscovich, K. Scholz, J. F. Hughes, and D. Salesin, "Customizable presentations," 2004 [Online]. Available: http://www.moscovich.net/tomer/papers/cpresentations.pdf.

[7] L. Good and B. B. Bederson, "Zoomable user interfaces as a medium for slide show presentations," *Information Visualization*, vol. 1, no. 1, pp. 35-49, 2002.

[8] S. Panjwani, A. Gupta, N. Samdaria, E. Cutrell, and K. Toyama, "Collage: a presentation tool for school teachers," in *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, London, UK, 2010, pp. 1-10.

[9] B. Perron and A. Stearns, "A review of a presentation technology: Prezi," *Research on Social Work Practice*, vol. 21, no. 3, pp. 376–377, 2011.

[10] D. Edge, J. Savage, and K. Yatani, "HyperSlides: dynamic presentation prototyping," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 2013, pp. 671-680.
[11] Y. Chen, M. Liu, J. Liu, Z. Shen, and W. Pan, "Slideshow: gesture-aware ppt presentation," in *Proceedings of 2011 IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, 2011, pp. 1-4.
[12] A. Fourney, M. Terry, and R. Mann, "Gesturing in the wild: understanding the effects and implications of gesture-based interaction for dynamic presentations," in *Proceedings of the 24th BCS Interaction Specialist Group Conference*, Swindon, UK, 2010, pp. 230-240.
[13] D. Glover, D. Miller, D. Averis, and V. Door, "The interactive whiteboard: a literature survey," *Technology, Pedagogy and Education*, vol. 14, no. 2, pp.155-170, 2005.
[14] X. Cao, E. Ofek, and D. Vronay, "Evaluation of alternative presentation control techniques," in *Proceedings of the Conference on Human Factors in Computing Systems (Extended Abstracts)*, Portland, OR, 2005, pp. 1248-1251.
[15] N. Negroponte, "The media room," in *Report for ONR and DARPA*. Cambridge, MA: MIT Architecture Machine Group, 1978.
[16] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: an overview," *Speech Communication*, vol. 57, pp. 209-232, 2014.
[17] S. S. Yurtsever, O. O. Cakmak, H. Y. Eser, S. Ertan, O. E. Demir-Lira, and T. Goksun, "Production and comprehension of co-speech gestures in Parkinson's disease," *Neuropsychologia*, vol. 163, article no. 108061, 2021. https://doi.org/10.1016/j.neuropsychologia.2021.108061
[18] G. Ali, M. Lee, and J. I. Hwang, "Automatic text-to-gesture rule generation for embodied conversational agents," *Computer Animation and Virtual Worlds*, vol. 31, no. 4-5, article no. e1944, 2020. https://doi.org/10.1002/cav.1944
[19] Y. Ferstl, M. Neff, and R. McDonnell, "ExpressGesture: expressive gesture generation from speech through database matching," *Computer Animation and Virtual Worlds*, vol. 32, no. 3-4, article no. e2016, 2021. https://doi.org/10.1002/cav.2016
[20] S. Harrison, "Showing as sense-making in oral presentations: the speech-gesture-slide interplay in TED talks by Professor Brian Cox," *Journal of English for Academic Purposes*, vol. 53, article no. 101002, 2021. https://doi.org/10.1016/j.jeap.2021.101002
[21] D. Schuler and A. Namioka, *Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.
[22] F. Kensing and J. Blomberg, "Participatory design: Issues and concerns," *Computer Supported Cooperative Work*, vol. 7, no. 3-4, pp. 167-185, 1998.
[23] P. Carayon, A. Wooldridge, P. Hoonakker, A. S. Hundt, and M. M. Kelly, "SEIPS 3.0: human-centered design of the patient journey for patient safety," *Applied Ergonomics*, vol. 84, article no. 103033, 2020. https://doi.org/10.1016/j.apergo.2019.103033
[24] M. Roccetti, C. Prandi, S. Mirri, and P. Salomoni, "Designing human-centric software artifacts with future users: a case study," *Human-centric Computing and Information Sciences*, vol. 10, article no. 8, 2020. https://doi.org/10.1186/s13673-020-0213-6
[25] G. Walsh, A. Druin, M. L. Guha, E. Foss, E. Golub, L. Hatley, E. Bonsignore, and S. Franckel, "Layered elaboration: a new technique for co-design with children," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, GA, 2010, pp. 1237-1240.
[26] H. Wu, W. Luo, N. Pan, S. Nan, Y. Deng, S. Fu, and L. Yang, "Understanding freehand gestures: a study of freehand gestural interaction for immersive VR shopping applications," *Human-centric Computing and Information Sciences*, vol. 9, article no. 43, 2019. https://doi.org/10.1186/s13673-019-0204-7
[27] A. S. Williams, J. Garcia, and F. Ortega, "Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3479-3489, 2020.
[28] N. Dahlback, A. Jonsson, and L. Ahrenberg, "Wizard of Oz studies: why and how," in *Proceedings of the 1st International Conference on Intelligent User Interfaces*, Orlando, FL,1993, pp. 193-200.
[29] B. J. Grosz, "The representation and use of focus in a system for understanding dialogs," in *Proceedings of International Joint Conferences on Artificial Intelligence*, Cambridge, MA, 1977, pp. 67-76.

[30] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through Wizard of Oz," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, Amsterdam, The Netherlands, 1993, pp. 277-284.

[31] J. D. Williams and S. Young, "Using Wizard-of-Oz simulations to bootstrap reinforcement-learning based dialog management systems," in *Proceedings of the 4th SIGDIAL Workshop of Discourse and Dialogue*, Sapporo, Japan, 2003, pp. 135-139.

[32] C. Munteanu and M. Boldea, "MDWOZ: a Wizard of Oz environment for dialog systems development," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000.

[33] B. G. Glaser and A. L. Strauss, and E. Strutzel, "The discovery of grounded theory; strategies for qualitative research," *Nursing Research*, vol. 17, no. 4, pp. 364-364, 1968.

[34] Z. S. H. Abad, S. D. Sims, A. Cheema, M. B. Nasir, and P. Harisinghani, "Learn more, pay less! lessons learned from applying the wizard-of-oz technique for exploring mobile app requirements," in *Proceedings of 2017 IEEE 25th international requirements engineering conference workshops (REW)*, Lisbon, Portugal, 2017, pp. 132-138.

[35] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "A grounded interaction protocol for explainable artificial intelligence," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, Montreal, Canada, 2019, pp. 1033-1041.

[36] C. Pagan Canovas, J. Valenzuela, D. Alcaraz Carrion, I. Olza, and M. Ramscar, "Quantifying the speech-gesture relation with massive multimodal datasets: informativity in time expressions," *PLOS One*, vol. 15, no. 6, article no. e0233892, 2020. https://doi.org/10.1371/journal.pone.0233892

[37] K. Ryselis, T. Petkus, T. Blazauskas, R. Maskeliūnas, and R. Damasevicius, "Multiple Kinect based system to monitor and analyze key performance indicators of physical training," *Human-Centric Computing and Information Sciences*, vol. 10, article no. 51, 2020. https://doi.org/10.1186/s13673-020-00256-4

[38] M. H. Heo and D. Kim, "Effect of augmented reality affordance on motor performance: in the sport climbing," *Human-centric Computing and Information Sciences*, vol. 11, article no. 40, 2021. https://doi.org/10.22967/HCIS.2021.11.040

[39] J. Nielsen, "Usability 101: introduction to usability," 2012 [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[40] S. Carbini, L. Delphin-Poulat, L. Perron, and J. E. Viallet, "From a Wizard of Oz experiment to a real time speech and gesture multimodal interface," *Signal Processing*, vol. 86, no. 12, pp. 3559-3577, 2006.