Evaluating Usability of Multiple Modalities in Mixed Reality

DT2140 - Multimodal interaction and interfaces 2020
KTH Royal Institute of Technology
Group 7
Supervised by Christopher Peters

Stefano Formicola formico@kth.se
Marcus Hamilton hamilt@kth.se
Charles-Eole Marichal marichal@kth.se
Siddharth Mishra sidmis@kth.se

Abstract

Presented here is a small puzzle game, built with Unity, which allows for multimodal input. This game is deployed on the Microsoft HoloLens to compare traditional input modalities, such as keyboards and controllers to modalities present in Mixed Reality (MR), such as gestures and speech. The performance of these are compared to previous studies and the results are presented using the System Usability Scale (SUS) which were questionnaires participants filled in, rating the usability of each modality. This is accompanied by their completion times, and their reflections from informal, unstructured interviews, to give a broader view. Haptics (Xbox controller and a joystick) scored highest in terms of the SUS and completion time, with keyboards as an ample substitute. Gestures performed poorly, and speech, only working in a limited setting, showed great potential. Due to limitations, generalizations about the usability of modalities in MR can not be made, but rather conclusions regarding the usability of modalities in an MR game. Evident from interviews with participants, some modalities inherently perform better in games. Improvements could be made with more scenarios, more participants, better selection using gestures and more processing power, or a better implementation for speech recognition to work properly.

1 Introduction

Background

When creating a new experience using new technologies, one must be careful to use the right modalities to set the interaction part. A problem that can easily occur is using a modality that is not appropriate to the task, thus damaging the user experience. The combination of keyboard and mouse to navigate 3D environments is one of the earliest used and most common modalities. In a study by Klochek and MacKenzie, a 3D environment is used to compare performance between an Xbox gamepad and a computer mouse in regards to tracking

the movement of an object [4]. Similarly, a study by Natapov et al. compared the Xbox gamepad, Wii remote and computer mouse to how users performed in pointing using Fitts' law [8]. When looking at new interaction technologies such as mixed reality, designers should wonder if such traditional controllers are still relevant.

While mixed reality has been shown to be more intuitive to show 3D data to achieve complex tasks [2], it has also been highlighted that augmented reality made simple puzzle-solving tasks simpler and more friendly than virtual reality for inexperienced users [7]. Moreover, using several modalities in augmented reality seems to improve the experience when performing some specific tasks [10]. Some have even tried to implement multimodal interactions in order to reduce cognitive load [6].

However, there are also constraints raised by those modalities. It seems that the adaptation from one platform to another is challenging when the user has already previous knowledge [3]. Also, even if those ways of interacting are more engaging, they seem to be more exhausting for users [9].

In this study, we want to explore the link between the use of multimodal interactions in augmented reality and the naturalness in the interactions. We work with the HoloLens, a device from Microsoft that allows users to use several modalities such as hand gestures, gaze or speech.

Purpose

We attempt to extend previous studies comparing input devices, such as mouse and controllers, by adding modalities such as gestures and gaze enabled by the Microsoft HoloLens. By creating a task within an interactive 3D environment in Unity, we measure usability of different modalities. We then evaluate users using the different modalities to solve the task and record their comments.

2 Project Design

Task

The main component of our project is the puzzle. It is a simple ball-in-a-maze puzzle which is rather calling for the user's dexterity than intelligence to solve it. We designed the maze with a very low complexity so that it can be solved quickly. It has globally the shape of a cuboid with a squared base and is seen from upside. A small transparent plane is on the cuboid, closing the box to prevent the ball from falling if the puzzle is put upside-down.

In order to solve the puzzle, the user has to tilt the box for the ball to roll inside until the ball has reached the end of the maze. Those are the only rules of the puzzle. Thus, the major medium used to solve the puzzle is gravity. Since most are quite familiar with this basic physical law, it does not require much previous knowledge to solve the puzzle.

The whole puzzle has been modelled and implemented with Unity 3D (v 2018.4.15.f1) and is displayed on the lenses of a Microsoft HoloLens device. This device is an augmented reality headset with advanced optics, multimodal sensors and holographic processing. The user has to wear the HoloLens to be able to see the puzzle. The output modality we use is then only vision which possesses high temporal and spatial resolution with a good representation of space, important characteristics to help the

user in their task.

The maze part of the box is rendered with a wooden texture while the transparent plane surface has been given glass reflectance properties. The idea is to make it look like a common real ball-in-a-maze puzzle so that a user with previous experience of this kind of puzzle immediately knows what he is supposed to do with a simple knowledge transfer. The wooden texture of the box also conveys the affordance that the puzzle can be moved just like a wooden box in real life. It is a commonly experienced material and the user can guess its properties such as its rigidity or weight by looking at it. The blue colour of the ball has been chosen for contrast purposes since the mean colour of the wooden texture is close to orange, the complementary colour is blue. All of those choices have been made to decrease the cognitive load induced by the puzzle understanding process.

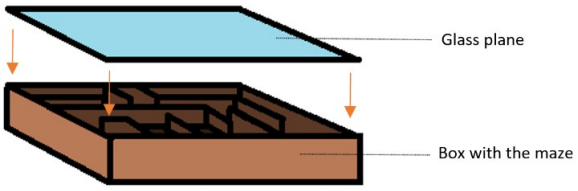
Modalities

The project is based on several input modalities to tilt the maze. No redundancy was introduced either in input or output since each input modality is tested separately from the other. However, if they were tested together, this would lead to some redundancy. The first modality is a joystick with both an Xbox controller and a joystick-only controller, the second one hand-gestures recognition from the HoloLens and the last one is speech recognition from the HoloLens too.

The controllers are quite simple to use. When tilting the joystick in one direction, the maze is also tilting in that direction. By the properties of haptic interfaces themselves, the joystick does not allow to have a great spatial accuracy when the user is not used to it. Keyboard is really basic and quite straightforward as the user can tilt the puzzle around one of the axes just by pressing a button. Although the accuracy is not ideal, the user can stop tilting very simply and is able to control several axes at the same time.

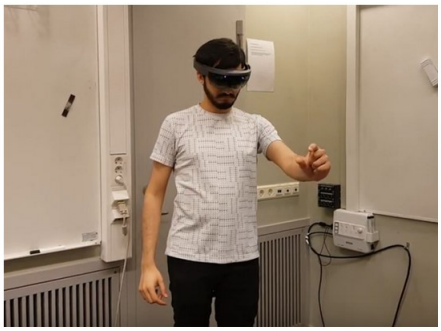
Hand-gesture recognition is made with the HoloLens framework [MRTK]. The maze is perceived by the HoloLens as a 3D object that can be either moved or tilted. Thus, eight small blue boxes are displayed around the puzzle basis: one for each cor-

(a) The 3D model of the maze

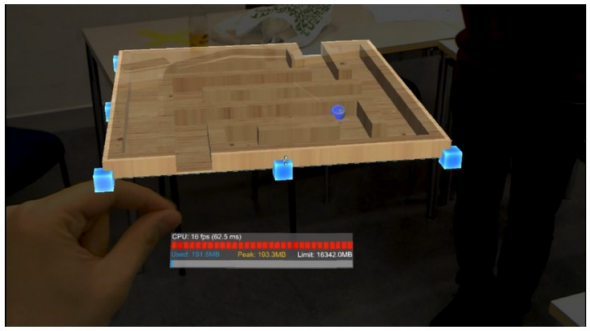


(b) Structure of the maze

Figure 1: Design of the maze



(a) Using hand-gestures



(b) User's point of view

Figure 2: Using the hand-gestures modality

ner and one at the middle of each side By doing a pinching motion while pointing at a blue box, the HoloLens allows the user to tilt or move the puzzle as long as the fingers of the user remain tight. It is necessary for a new user to learn that closing their hand on a blue box allows that kind of movement. A short time of adaptation can be required then. The interesting counterpart is that the user does not have to wear any marker to be able to interact with the system.

The HoloLens MRTK for Unity3D features a section where Speech Input can be enabled. Enabling this section gives users options to add custom words called Keywords and assign some action to it. For the HoloMaze, we chose Keywords like "Turn Up", "Turn Right", "Turn Left" and "Turn Down". In the maze 3D component, we assigned a Speech handler to which we attached scripts to manipulate the maze orientation whenever the keywords are spoken.

Since the puzzle is not real, the user cannot hold the maze. Contour-following and enclosures around the box are impossible, making the understanding of the shape harder. They also do not have any sense of weight and cannot feel the vibrations of the ball rolling down, complicating the task.

A demo of the application can be found following **this link** [1].

3 Evaluation

To test the multimodal designed system we decided to focus our attention only on a small sample of participants. This is due to how strongly the project has been affected by the limitations and difficulties in restrictions regarding gatherings of people during the COVID-19 pandemic. This should therefore be considered a pilot study of the usability evaluation of a multimodal mixed reality application. If an A/B testing would have required a lot of participants to test the different input modalities, according to Shneiderman et al. in their study from 2016 the questionnaire evaluation method allowed the execution of a small but meaningful study [11]. We considered the well known System Usability Scale developed by Brooke et al. in 1996 the most suitable questionnaire to evaluate our system [1], based on ten statements with which users rate their agreement (see Appendix).

[1] `https://vimeo.com/498943851`

Method

The test takes place with different settings, given in a randomized order to improve the validity of collected data. Tasks differ from each other only in the input modality used in combination to the mixed reality application (hand gestures, controllers, keyboard), that is a HoloLens 3D projected maze. Users are required to complete the task by finishing the game, which consists in tilting the maze over the 3 axes to move a ball from the start point to the end.

Before discussing with participants about the experiment they have been asked to fill in the questionnaire with all items to be checked and they have been told to, if unsure about a specific statement, mark the centre point of the scale. The overall usability is calculated as a weighted sum of the individual items. Odd numbered items' contribution is the scale position minus 1 and even numbered items' contribution is 5 minus the scale position. This is due to the fact that positive and negative items are alternated in order to prevent response biases and make users make an effort to think about their answer as highlighted by Brooke in 1996 [1]. At the end, the sum of the weighted scores is multiplied by 2.5 to obtain the overall value of System Usability, on a scale 0 - 100.

Results

Given the limitations described above, we managed to conduct our experiments on 5 people, all of them students of a technical university in Sweden. Even though the results of both controllers (the joystick and Xbox controller) are very similar, we decided for practical reasons to leave them separated in the data analysis, due to the fact that the experimentation phase started before receiving the suggestion to only use one of these two modalities by our supervisor.

We analysed the SUS results with the formula described in the Methods section and obtained the following results, plotted and described in the table on figure 3.

As it is possible to notice, the Xbox controller and Joystick modalities performed best and are the preferred modalities in terms of usability, with an average score of, respectively, 89.5 and 89. The gesture-based interaction, instead, performed worse.

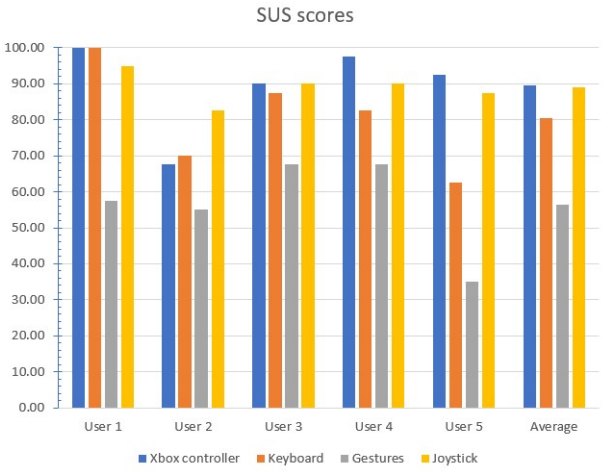
These results are coherent with the completion times collected during each task of the experiments. The preferred input modalities, indeed, are also the ones that helped users to complete the game in a shorter time, while the gesture-based interaction made it difficult for some users to finish the game in a reasonable amount of time.

4 Discussion

Due to certain limitations such as restrictions and guidelines regarding COVID-19 and that this was conducted partially during the winter holidays, the number of participants was limited. The limited number of participants testing our application was anticipated, meaning the lack of quantitative data was attempted to be complemented by more qualitative data. The participants were in unstructured interviews informally asked about the different modalities and how they responded to the application. The thoughts, feelings and experiences the participants had while testing this application are addressed in this section. A SUS fails in expressing deeper meanings and

	Xbox controller	Keyboard	Gestures	Joystick
User 1	100	100	57.5	95
User 2	67.5	70	55	82.5
User 3	90	87.5	67.5	90
User 4	97.5	82.5	67.5	90
User 5	92.5	62.5	35	87.5
AVG	89.5	80.5	56.5	89
SD	12.91	14.72	13.29	4.54

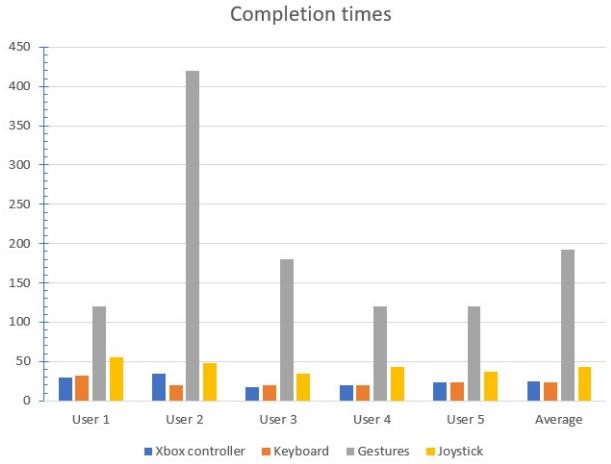
(a) SUS scores



(b) SUS scores histogram

	Xbox controller	Keyboard	Gestures	Joystick
User 1	30s	32s	120s	55s
User 2	35s	20s	420s	48s
User 3	17s	20s	180s	35s
User 4	20s	20s	120s	43s
User 5	23s	23s	120s	37s
AVG	25s	23s	192s	43.6s
SD	7.38	5.20	130.08	8.17

(c) Completion times



(d) Completion times histogram

Figure 3: Results from the study

reflections and this complements what the SUS and results lack.

Participants were not able to test the speech modality, due to reasons explained in the Limitations section. Speech will thus not be in focus in this section but is discussed more in the following sections. The project supervisor expressed that comparisons between speech and gestures would be the most valuable for a project such as this, which is why those sections place much focus on speech, despite its limitations.

All users performed significantly better in regards to completion time using the controllers rather than gestures. The Xbox controller specifically performed better than any other input. This could be because users were generally more experienced using controllers, compared to gestures. The superior performance of the Xbox controller could have other implications. Since what was created in this project was a recreation of a traditional game, existing both physically and virtually, (and now holographically) it is implied that for games, controllers are more appropriate, which Pirker et al. also suggest [9]. It does not imply that gestures perform worse at interacting with holographic components. This is reinforced by what the users expressed. They agreed

that using a controller felt most natural out of all modalities in interacting with the tilting of the board and knew exactly how to manipulate the board to move the ball in the desired direction. Unlike when using gestures they expressed uncertainty of how to interact with the maze, despite given prior instructions. However, most expressed that they felt gestures have potential and could either be better utilized in other scenarios, or in a better implementation. To produce more conclusive results regarding MR in general, rather than about a specific type of application, one would have to extend the study by creating multiple scenarios. Both including and excluding games. These would have to be tested on a larger set of participants, preferably in randomized order to avoid biases. Compared to keyboards, the reason controllers performed better is likely due to precision. Controllers accept gradual input, unlike keyboards which are binary, full or none. This gives controllers better precision and what users expressed as the right "feel".

Further, the implementation and handling of gestures have large consequences on the results. This particular implementation requires users to manipulate the board using a pinching gesture on a cube located in the centre of each side of the maze (see fig. 2). The size of the cubes or the precision of the gesture recognition by the HoloLens were considered too unforgiving. The pinches were not always recognized, which might be a result of the limitations of the HoloLens. To perform the gestures users had to have the cube selected using the gaze cursor, looking directly at it. The HoloLens does not track eye movement, leaving the users limited to a cursor centred in their field of view. Had the cube been larger this might have been less of an issue, but the users also felt that they wanted to observe the board while tilting it. By forcing them to gaze directly at the cube this was rendered more difficult and could contribute to why gestures performed worse than anticipated. Improvements to the results could be made by implementing eye-tracking in later generations of the HoloLens to allow for more freedom of movement. Further improvements could be made if the manipulation of objects were not limited to adjustment cube, but rather customizable to cover the entire area of the sides of the maze.

Reflection

This section contains the reflections of the authors and our experience with the HoloLens. Unlike the other participants, we were able to test speech as an input modality, although in a limited setting. We also had more time to familiarize with the HoloLens, exploring its different features and navigating its menus.

In contrast to the results, our hypothesis prior to creating the maze was that gestures would outperform any other modality. This was based on the initial interactions with the device. Navigating menus using gestures was intuitive and efficient. Although there was no support for controllers, some menus could hypothetically be navigated using them. This would be advantageous when the arms start to tire after being held up for longer periods of time. However, moving items, such as a browser in three dimensions seemed inferior to anything other than gestures. This led to our assumption.

In the initial stages of the project, it was theorized gestures would perform better than speech too. Unless the commands

were known it was quite challenging navigating using speech only. However, having tested speech on the maze, it was quickly realized that this would perform significantly better than gestures. The voice commands were simple and obvious, such as "Go up" or "Up". It eliminated the issue gestures had with precision in pinching the adjustment cubes. The user was not limited by the gaze cursor either. Removing these components saved both time and frustration. Commands could be given in any situation regardless of the position of the user, and unlike using gestures the user did not suffer exhausted arms.

In comparison to the controllers, particularly the Xbox controller, favoured by all participants, speech did not limit movement. The controllers require to be plugged in, limiting the user's movement to the length of the controller's cord. This leads us to believe that speech is potentially the best-suited modality for applications of similar types that require more movement. Further, we believe that if we were less limited by time-constraints, or had more processing power we could have an implementation where speech performed on par with, or possibly better than controllers. These speculations are further reinforced by similar studies. A study from 2013 by Lee et al. comparing speech against gestures and a fusion of the two in an AR environment showed at ($p < .01$) the completion time with the gesture interface was significantly different (slower) from speech [5] .

Another explanation is that gestures are not well-suited for gaming, as previously discussed. A study comparing keyboard controls and gestures in gaming show that the SUS of keyboards is not only significantly higher (20 points), but gestures also perform below average according to the SUS rating methodology [9]. The average lies on 68 points, whereas the gestures scored 55. Compared to our testing where gestures scored 56,5 we see a clear tendency in games.

Limitations

The most pressing and most obvious limitation of our pilot study is the small sample size. To reach higher levels of confidence, especially regarding any possible differences in modalities between the tested conditions, more participants would need to be recruited.

This would likely also address another major issue regarding our sample: the lack of diversity. Due to our reliance on convenience sampling during a time where contact outside of one's immediate group of acquaintances is rather difficult or even prohibited, we were unable to recruit any participants outside the age range of students. For a full-scale study, more emphasis should be put on gathering a sample representing a larger part of the potential user group of MR applications.

Another major limitation we faced while developing this project was Microsoft's out-of-date drivers for HoloLens 1, whose development has been stopped in favour of HoloLens 2. This made us relying on legacy SDKs and older versions of Unity, as well as old posts on the web when needed.

Another limitation lies within the hardware, which has limited the number of modalities we could test. Specifically speech. Speech proved incredibly unreliable in our application. Most commands were not recognized at all or took several attempts to be recognized. Since speech

recognition is handled internally by the HoloLens, this was nothing we could affect. When it performed poorly the strain on the processing was highly noticeable. The frame rate was significantly low, oscillating between 15-20 at worst and 30-40 at best. Meanwhile, the CPU usage was high, often reaching 100% usage.

However, a discovery was made due to network connectivity issues. It caused the textures to barely be rendered, making it pixelated and virtually impossible to distinguish the features of the maze. In turn, the performance in regards to processing and frame rate went up. At this stage, the speech recognition worked flawlessly, but the rendering quality was too poor to use in user tests, as they would affect the overall user experience.

The project also suffered from limited time with the equipment. The HoloLens had to be borrowed, which limited use to be on campus. This meant relying on free schedules of three parts, the person who would lend the equipment, at least one person in the group to borrow the equipment, and participants to test the application. This, in combination with winter holidays and the pandemic, meant that finding available time-slots that worked for all three parts was challenging, and in some cases not possible. This further restricted the number of participants.

5 Conclusion & Future work

With the limitations, reflections and discussions in mind, two details need to be addressed. First, this is and should be considered a pilot study. Second, generalizations about modalities in MR cannot be made. It has, however, given insights in the usability of modalities in small physics-based MR games. We have also gathered insights on how to improve and extend this to conduct a study that could yield more conclusive results with more confidence.

Manipulating objects, in terms of tilting a board by moving its sides up or down, haptics had the best performance. Users felt it was natural, intuitive, easy to maneuver and most suitable for the task. Keyboards performed as a suitable substitute, also reflecting results of Gerling et al. but users expressed it lacked the finesse and feel of the haptics. Comparing these to gestures showed that gestures might not be suitable for games, but show potential in other areas. Speech input seems to have the most potential, as it performed well in this scenario, as well as interacting with other items and menus in the HoloLens. It also did not restrict users to an area with limited range and was able to solve the task quickly and efficiently. In the study from 2013 by Lee, Billinghurst, Baek, et al. speech is shown to perform better than gestures only, and multimodal fusion of the two is suggested [9] and as it is also suggested by Piumsomboon et al. [5]. We would therefore make the same suggestion.

To improve upon this in the future, certain aspects need to be taken into consideration from what was learned here. Firstly, there need to be different tasks to be able to generalize about the usability of modalities in MR. Here it was limited to a game, where certain modalities inherently perform better than others. Further, no fusions of modalities were used. Particularly, a fusion of gesture and speech could enhance performance and overall user experience. An-

other important aspect of the modalities explored here is movement. This task did not require users to move. Movement would better utilize the wearability of the equipment, and hence tasks that naturally require movement would make for a better comparison between corded haptics to speech and gestures. Eye-tracking for better gesture recognition and higher processing power for the issues with speech recognition is also suggested for future generations of hardware, as they would improve the interactions of these modalities. Gestures would also benefit from having larger areas that allow for manipulation, rather than adjustment cubes. These areas should be customizable to be able to cover, for example, the side of the maze. Noticeable is also the similarity in performance between the joystick and Xbox controller. One should consider discarding one of these in future tests, as the order of testing might affect the performance, due to the similarity in the nature of operating them.

References

[1] Brooke, J. (1996) SUS: A "Quick and Dirty" Usability Scale. In *Usability Evaluation In Industry, by P. W. Jordan, B. Thomas, B. Weerdmeester, and I. L. McClelland.*

[2] Cartucho, J., Shapira, D., Ashrafian, H., & Giannarou, S. (2020). Multimodal mixed reality visualisation for intraoperative surgical guidance. *International journal of computer assisted radiology and surgery*, 15(5), 819–826. doi:10.1007/s11548-020-02165-4

[3] Gerling, K.M., Klauser, M., & Niesenhaus, J. (2011). Measuring the impact of game controllers on player experience in FPS games. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (MindTrek '11). Association for Computing Machinery, New York, NY, USA, 83–86. doi:10.1145/2181037.2181052

[4] Klochek, C., & MacKenzie, I. (2006). Performance Measures of Game Controllers in a Three-dimensional Environment. *Proc Graph Interface 2006*. Toronto, Ont., Canada, Canada: Canadian Information Processing Society. 73-79. doi:10.1145/1143079.1143092.

[5] Lee, M., Billinghurst, M., Baek, W. et al. A usability study of multimodal input in an augmented reality environment. *Virtual Reality 17*, 293–305 (2013). doi:10.1007/s10055-013-0230-0

[6] Link S. et al., (2016) An intelligent multimodal mixed reality real-time strategy game, *IEEE Virtual Reality (VR)*, Greenville, SC, 2016, pp. 223-224, doi: doi:10.1109/VR.2016.7504734.

[7] Moon, M., & Kwon, C. (2019). Developing a Puzzle using the Mixed Reality Technology for the Elderly with Mild Cognitive Impairment. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075,Volume-8 Issue-8S2, June 2019 `https://www.ijitee.org/wp-content/uploads/papers/v8i8s2/H11500688S219.pdf`, retrieved on 29/11/2020

[8] Natapov, D., Castellucci, S.J., & MacKenzie, I.S., (2009). ISO 9241-9 evaluation of video game controllers. In *Proceedings of Graphics Interface 2009* (GI

'09). Canadian Information Processing Society, CAN, 223–230.

[9] Pirker, J., Pojer, M., Holzinger, A., Gütl, C. (2017) Gesture-Based Interactions in Video Games with the Leap Motion Controller. In: *Kurosu M. (eds) Human-Computer Interaction. User Interface Design, Development and Multimodality. HCI 2017.* Lecture Notes in Computer Science, vol 10271. Springer, Cham. doi:10.1007/978-3-319-58071-5_47

[10] Piumsomboon, T., Altimira, D., Kim, H., Clark, A., Lee, G. & Billinghurst, M., (2014). Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality, *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Munich, 2014, pp. 73-82, doi: doi:10.1109/ISMAR.2014.6948411.

[11] Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N., (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* 6th ed. Pearson.

Appendix

System Usability Scale (SUS)

To evaluate the multiple modalities we asked participants to answer a single SUS questionnaire modified for each input modality.
The general template of the SUS we edited is reported below and users are asked to choose an option on a range between 1 and 5 where 1 indicates "strongly disagree" with the statement and 5 means "strongly agree".

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.