Multimodal Interface Study of Voice and Modified Gestures on Immersion and Effectiveness

ISABELLA BARNARD*, JOHN MAKSUTA*, and JADEN MASCARENHAS*, Colorado State University, USA

An introduction to Human Computer Interfaces and Virtual Reality. Related works are discussed and provide some context for the topic. Our methodology is then discussed and provided a detail the study we are conducting.

CCS Concepts: • **Human-centered computing → Gestural input**; **Pointing devices**; **Sound-based input / output**; **Empirical studies in HCI**; **Virtual reality**.

Additional Key Words and Phrases: Multi-modal

ACM Reference Format:
Isabella Barnard, John Maksuta, and Jaden Mascarenhas. 2018. Multimodal Interface Study of Voice and Modified Gestures on Immersion and Effectiveness. 1, 1 (March 2018), 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Human Computer Interfaces and Virtual Reality are great technologies for bringing users into environments that are often easier to produce that building or producing the real thing. There are many uses for Virtual Reality, and the input methods are often very significant and can vary greatly. The way a user interacts with a device is called a modality, and we will be evaluating a multi-modal interface where the user will be using two different interfaces to perform certain tasks.

For our project, we are going to create a 'spell casting' game. The focal idea for our project is to use speech as our main form of modality and to test various modalities against this to determine which is the most comfortable, easily utilized, and immersive for the user. We will be testing whether users prefer hand gestures, controllers, or the Logtitech MX Ink Combo pen, in addition to speech, to cast spells. To measure the user's ease-of-use, we will gather information regarding the response time and other quantitative data from our test group. From here we will be able to determine which modality should be used in conjunction with speech to create the optimal experience of the game and fine-motor VR experiences overall.

2 Related Work

Studies have been performed regarding some of the issues users may have while drawing in a VR space. In the study, the researchers addressed the fact that 2D projections often allow the user to communicate easily, while the 3D space is more difficult to utilize. The issue arises in response to a lack of understanding related to depth perception in a 3D space. It is discussed that "The primary challenge to 3-D drawing in-air is the accuracy of

*all authors contributed equally to this research.

Authors' Contact Information: Isabella Barnard, isabelba@colostate.edu; John Maksuta, jmaksuta@colostate.edu; Jaden Mascarenhas, c835223467@colostate.edu, Colorado State University, Fort Collins, Colorado, USA.

ACM XXXX-XXXX/2018/3-ART
https://doi.org/XXXXXXX.XXXXXXX

users' strokes". Accuracy tends to decline while the user utilizes VR, due to the process often being more complex in nature than their 2-D counterparts[9].

Besides depth perception, structured sketches are also difficult to complete in VR because of low precision. "The low precision of 3D sketching makes it difficult for users to accurately express their drawing intentions". Due to a lack of accurate expression, many users may opt out of using VR sketching tools. A visual guidance tool may improve precision and depth perception. A guidance tool may use continuous stroke recognition and guide lines to improve user experience. The study concluded that using guide tools helped improve structural accuracy[4].

Despite these issues, many VR applications have been created for artistic expression. These applications include drawing, modeling, and animation applications. By using auto-correction, many applications are able to tidy up lines, allowing the user to have a more controlled experience. In order to do this, the user needed access to canvas manipulation. The goal of the study was to "enable users to sketch freehand in 3D with improved precision while maintaining explicit user control". In the study, completion time and drawing precision were the main aspects being tested. The participants were asked to draw circles, straight lines, bows, tildes, and eclipses. In the study the shapes "are designed to be drawn in a single, continuous motion". In the study, it was concluded that moving the canvas contributed to an "improvement in both precision and smoothness", which had the greatest impact on lowering completion time[8].

A benefit of using VR for drawing and creative expression is that it is not as restrictive as real-life drawing. The canvases are infinite, and the user can create almost anything they want. In the study, it is discussed that "VR allows for expression that is unrestricted by natural physical laws". This can be extremely appealing for those who want to create an environment in VR that is not practical in real life. In our case, this would be related to our game as w hole. We can create an environment for the user in which casting spells is possible[3].

In terms of multi-modal interaction, a study was done regarding how combining hand gestures and drawing may lead to more accurate sketches. The researchers used a "formative study to compare two-handed interaction with tablet-pen interaction for VR sketching". They used gesture capture to use one hand for a canvas while the other created the drawings with one finger. The researchers faced challenges regarding a limited region, fingertip sensitivity, uneven canvases, and instability. Using a canvas for a hand caused a :limited workspace for drawing". It was difficult to track the fingertips because the gloves relied on "tensions on the bending joints". The uneven canvas was caused because the hand is "not as flat as a canvas". Instability was caused because "hands are neither fixed nor rigid". Despite these setbacks, the researchers allowed "users to pan the canvas by using a pinch gesture". They used a "calibration step" to calculate the fingertip distance. To fix the uneven surface, they needed to "separate the hand into several sub-regions". Lastly, they used algorithms to smooth the strokes. This system improved the accuracy and usability of the system by fixing the problem of "how to beautify strokes without reducing aesthetic quality and affecting users' intentions"[10].

For symbol recognition in VR spell-casting games, a study was done regarding line-based gesture recognition. In this study the researchers used a recognition program in order to "transform user input into an ordered list of vectors between each input point". This process was reliant on a database, which helped recognize the shapes. They tested this game mechanic with their research group, and found that "spell casting had mixed results". The process of learning how to cast the spells was easy, while the complexity of the symbols seemed to create issues. This difficulty that their users encountered was likely due to a lack of "verbal or visual demonstration" they stated in their findings. A proposed solution for this problem in their study was to increase the "number of symbols that can be recognized" and emphasized "a need for better symbol recognition methods"[5].

Similar studies have been performed that examined the use of gestures in a One- or Two-Handed modality, using task based comparisons of the user interfaces[7]. In their study, they wanted to find out about the "adoption process of VR as a tool" and were analyzing two hand-tracking modalities, these being One- and Two-Handed[7].

In their analysis they evaluated the "effectiveness of the chosen gestures and how they suit either interface"[7]. This is similar to the approach that we are taking in evaluation of multi-modal interactions, although our scope is more limited. This study concluded that the One-Handed user interface was preferred over the Two-Handed interface except for the group who tested the One-Handed interface first. Their study concluded also that One-Handed interfaces were faster to complete tasks on average, and that right handed gestures were preferred, although the vast majority of participants were right-handed[7]. Our study is focused on certain quantitative metrics and task completion time is one of them.

Input methods are often evaluated with many different human computer interactive technologies. For the head-worn displays, such as the Oculus Quest, the input methods most explored were freehand interaction, followed by head-based input, while "hardware-based and speech interaction were considered the least, but still occurred relatively frequently" [6]. Our study will evaluate a variety of inputs, each used as multi-modal input with a speech interaction. The multi-modal input combinations most frequently evaluated, in a 2023 review of interaction techniques, were head and hardware controller, followed by hand and speech, and hand with head[6]. We will be using the hand and speech multi-modal interaction techniques.

Speech input methods had many interesting findings in the 2023 review. Although, it was found to be less natural for single user applications and more appropriate for collaborative environments with multiple users. In educational applications, voice input was found to "increase memory retention and learning" [6].

The number of tasks performed for evaluation for the head-worn displays varied from rarely one or four, and the most two or three tasks. Only one study in the review evaluated voice and hand gestures for a task that interacted with a virtual character[6].

Another study by Barry A. Po, et al. used the "two visual systems hypothesis" which provides information about the "complex relationship between visual perception and human motor movement". This study similar to ours explored the multi-modal interfaces of voice and gesture, however in their study, they were concerned with "spatial target selection" and made several predictions which their study experiments were conducted to test. The results of their study demonstrated how "the presence or absence of certain visual cues... tracking cursors and asymmetric frames, can influence voice and gestural interaction for spacial target selection" [1]

Although this study utilized the same multi-modal interfaces for interaction in the virtual environment, their use was for target selection, where ours is utilized for task completion, in conjunction.

Voice and gesture interaction has also been used in the teaching scenarios, and the "implementation and design of gesture and voice-based interaction in virtual teaching scenarios" was explored in a study by Ke Fang and Jing Wang[2]. They used voice recognition and gestures to complete a series of tasks including teleportation, adjusting lighting levels, and entering numbers in a simulated keypad. This was compared against completing the tasks with controllers only. In the experiment, tasks 1 and 2 were longer and task 3 was shorter using voice and gesture versus controller. Although it was not entirely conclusive due to the scope whether voice and gesture are better inputs than controller, it demonstrated "enhancing the interactive experience in virtual teaching" [2].

3 Methodology

Using an Oculus Quest 3 and a virtual reality program environment written in Unity, we will evaluate the use of voice commands and gesture input. The gesture inputs come from three different modalities that are under evaluation: controllers, hands, and pen device. Voluntary subjects will be recruited and information will be provided to obtain informed consent from the participants in the study. Subjects will be given a randomized order of the three variations of modalities, Speech and Controllers, Speech and Hands, and Speech and Pen device. We will utilize latin squares to randomly assign the order of the tests for each participant. The Control variable in this study is the speech input which will remain constant and be present in each test. The independent variables are the second modalities.

Each task will require the user to "cast a spell" of some type and either target an object or have it automatically target the object in the game. The spell will be considered successfully cast when the user successfully completes each modality task within a certain time frame, such as 5 seconds. The speech modality is successful when the user says an acceptable phrase, such as "bipity bopity boo" or "abra kadabra", etc. The independent variable modalities will be considered successful when the user draws a certain symbol path using their gesture input. If the user completes these two tasks near simultaneously, the task is considered successful, and the user will see notification of the success in the form of the spell being cast, whatever that particular spell looks like. The failed tasks will result in a different notification where the spell "fizzles" and a certain animation, sound, or effect will be displayed to notify the user of the failure. The number of successes and failures will be counted and recorded, as well as the time it takes to complete each part of the task and the entire task itself. After completion of each independent variable modality, the user will be presented with the end of the game and either complete a survey on the device in the game, or complete a survey that is distributed to them by the researcher.

The results of the qualitative and quantitative data stored will be analyzed and compared to answer the following questions and each modality will be ranked for each question.

- `Ease of use`: Which modality in conjunction with speech was easiest (which took least time)?
- `Preference`: Which modality did the participant prefer?
- `Effectiveness`: Which modality was the most effective?

3.1 Hardware

The hardware devices are the Quest 3 headset, the included pair of controllers, and the Logitech MX Ink Combo Pen. The quest 3 headset uses a series of cameras to determine the position of hands when using the hand inputs, rather that the controller. It uses the head mounted display to display the virtual environment to the user. The display is capable of using its cameras to display and Augmented Reality mode called pass-through, where the cameras display their output in real-time to the head-mounted display through which the user is viewing. This creates the perception that one is seeing their real-world surroundings while also allowing placement of virtual objects in the environment. For our purposes, we will be using the full virtual reality display without the pass-through feature, but it is worth noting that the multi-modal inputs of voice and gesture could have many use cases in Augmented Reality as well.

3.2 Evaluation

Subjects will be assigned to complete tasks in an order determined by latin squares to randomize the order of tests. There are three tasks the subject will complete, voice and controllers, voice and hands, and voice and pen. Each multi-modal pair will involve the user saying a command phrase and performing a specific gesture within a given time frame. When both tasks are successful and the time is within a certain limit, the task will be completed successfully. The number of attempts, and completion time will be recorded for analysis.

In the effort to collect qualitative data we will also have a survey that will either be distributed after the subject completes the tasks, or it will be displayed as a software user interface for the user to complete at the end of the tasks. The data will elicit the opinions of users on a scale of various data points such as, ease of use, immersion, realism, and preferences. We will also collect some demographics information such as age and prior experience with XR technology.

References

[1] Brian D Fisher Barry A. PO and Kellogg S Booth. 2005. A Two Visual Systems Approach to Understanding Voice and Gestural Interaction: Lanquage, Speech and Gesture for VR. *Virtual reality: the journal of the Virtual Reality Society* 8, 4 (2005), 231–241. doi:10.1007/s10055-005-0156-2

[2] Ke Fang and Jing Wang. 2024. Interactive Design with Gesture and Voice Recognition in Virtual Teaching Environments. *IEEE access* 12 (2024). doi:10.1109/ACCESS.2023.3348846

[3] Christian Jones Janice Tan, Lee Kannis-Dymand. 2023. Examining the potential of VR program Tilt Brush in reducing anxiety. *Virtual Reality* 27, 4 (Dec 2023), 3379–3391. doi:10.1007/s10055-022-00711-w

[4] Shuxia Wang Weiping He Jingjing Kang, Shouxia Wang. 2021. Fluid3DGuides: A Technique for Structured 3D Drawing in VR. *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Dec 2021), 1–2. doi:10.1145/3489849.3489955

[5] L. Quirk J.J. LaViola O.C. Jenkins M. Katzourin, D. Ignatoff. 2006. Swordplay: Innovating Game Development through VR. *IEEE Computer Graphics and Applications* 26, 6 (Nov 2006), 15–19. doi:10.1109/mcg.2006.137

[6] Frutos-Pascual M. Creed C. Williams I Spittle, B. 2023. A Review of Interaction Techniques for Immersive Environments. *IEEE Transactions on Visualization and Computer Graphics* 29, 9 (2023), 3900–3921. doi:10.1109/TVCG.2022.3174805

[7] Teijo Lehtonen Taneli Nyyssönen, Seppo Helle and Jouni Smed. 2024. A Comparison of One- and Two-Handed Gesture User Interfaces in Virtual Reality–A Task-Based Approach. *Multimodal technologies and interaction* 8, 2 (2024), 10–. doi:10.3390/mti8020010 web.

[8] Can Liu Mingming Fan Tianren Luo Zitao Liu Mi Tian Teng Han Feng Tian Xiaohui Tan, Zhenxuan He. 2024. WieldingCanvas: Interactive Sketch Canvases for Freehand Drawing in VR. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (May 2024), 1–16. doi:10.1145/3613904.3642047

[9] Akshay Sharma Yotam Gingold Xue Yu, Stephen DiVerdi. 2021. ScaffoldSketch: Accurate Industrial Design Drawing in VR. *The 34th Annual ACM Symposium on User Interface Software and Technology* (Oct 2021), 372–384. doi:10.1145/3472749.3474756

[10] Hongbo Fu Alberto Cannavò Fabrizio Lamberti Henry Y K Lau Wenping Wang Ying Jiang, Congyi Zhang. 2021. HandPainter - 3D Sketching in VR with Hand-based Physical Proxy. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (May 2021), 1–13. doi:10.1145/3411764.3445302

A Online Resources

Our code repository is located on GitHub and is available here:
https://github.com/csu-hci-projects/CS465$_{M}ascarenhasJ_{B}arnardI_{M}aksutaJ$