The Efficacy of Voice Activated Controls

CAMDEN MCGINNIS, JUSTIN GOING, KIMERON LAZARE, and ANDREW JELKIN, Colorado State University, USA

The purpose of this research is to establish a baseline understanding of the benefits and limitations of sound-based input as a modality. This work presents a modality study in which speech interaction is directly compared to pointer-based input using the Simon Electronic Memory Game as a vessel. Twenty participants from Colorado State University were recruited to participate in the study. Each participant was asked to complete a game using each of the modalities and their inputs were closely analyzed to compare both factors. The data showed that... (TBD)

CCS Concepts: • **Human-centered computing** → **Sound-based input / output**; *Pointing devices*; Displays and imagers.

Additional Key Words and Phrases: gesture and speech interaction, speech recognition

ACM Reference Format:
Camden McGinnis, Justin Going, Kimeron Lazare, and Andrew Jelkin. 2023. The Efficacy of Voice Activated Controls. *J. ACM* 37, 4, Article 111 (August 2023), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Sound based inputs have existed as a novelty since the 1950s, but did not see wide use until Google voice search in the early 2000s [8]. Now, there's voice control in vehicles, appliances and other electronic devices. One particular area of interest with limited voice input features is the field of video gaming. The market size of the video game industry was approximately 220.79 billion in 2022 [7]. Like virtual reality, voice activated control in gaming offers a more inclusive modality that has the potential to significantly enhance the experience. The purpose of this paper is to compare sound-based input with the more traditional pointer-based input to assess the efficacy of the different of modalities. This paper reports the results of an experiment in which voice activated control was introduced into an existing game environment.

2 PRIOR WORKS

2.1 The Vocal Joystick

The vocal joystick engine maps a pointer to human vocal signals as an alternative to traditional mouse pointer. The system uses voice pitch, quality and energy, which gives four degrees of freedom. [3] The VJ engine was tested against eye tracker inputs, and produced similar results.

Authors' address: Camden McGinnis, camden.mcginnis@colostate.edu; Justin Going, justin.going@colostate.edu; Kimeron Lazare, kimeron.lazare@colostate.edu; Andrew Jelkin, andrew.jelkin@colostate.edu, Colorado State University, Fort Collins, Colorado, USA.

Several improvements to the Vocal Joystick engine have been made since inception. The Voice Game Controller expands on the base mouse control of the original joystick by implementing a Vocal D-pad for gaming, as well as implementing more natural language processing strategies. [6]

Sound-based input is more natural for the user to perform, making the interaction between human and computer much more simple. [12] This is likely because the commands are much more literal, and thus easier to comprehend.

Through numerous studies, it has been shown that voice controlled mouse systems such as the Vocal Joystick are user friendly, user independent, cheap, and accurate. [9] These factors show that sound-based input modalities have the potential to be greatly useful for general purposes.

2.2 Applications of Sound-based Input

There exists some web browsers such as Conversa which allow the user to navigate the web using entirely vocal commands. The use case for a browser such as this is to help those with motor impairments to easily browse the web. [5]

Sound-based inputs are on the rise, and have the potential to make certain systems more streamlined and easy to use. Some more examples of applications include voice systems in cars so that the driver can micro manage without taking their eyes off the road, [4] monitoring systems in hospitals to provide faster emergency care to heart-failure patients, [10], or create an interface for life-log retrieval to streamline the process [1]

2.3 Other Speech Command Implementations

Gaze-voice control combines gaze control (eye tracking) with voice commands to elicit a completely hands-free gaming experience. While it was found that gaze-voice control was less easy to use than touch inputs, users who played with gaze-voice reported being significantly more immersed with the gaming experience than those who use touch inputs. [11]

The purpose of this research is to analyze the strengths and weaknesses between voice control and mouse control. These prior works help give us some idea of the challenges and applications that voice control can have in the field. These idea will then be applied to the world of gaming, where we hope to gain a better understanding of how voice control can work in this paradigm.

In the future, we hope this research will help move speech command into the larger computing paradigm, so that it may become a ubiquotous part of computing .[2]

3 METHODOLOGY

We conducted a study comparing the use of voice commands against traditional mouse inputs. We investigated the difference in performance of each input modality through an application we designed.

3.1 Application

The application is a web-accessible page consisting of front-end, middleware, and database with a stack consisting of Angular, Fast API, and Postgres. The stack built on python and typescript is containerized by Docker and hosted on the web. Upon their first visit to the site, users are presented with a login screen prompting them to create an account. Users must create an account so that they can be assigned to a group and have their data stored on the application server. When creating their account, the user is prompted to fill out demographic information about themselves, after which they can begin the experiment. After logging in, the user is presented with the start screen, which provides a

tutorial for playing the game and a button allowing them to start the game. The game that the user can play is Simon Says. Participants are tasked to memorize the color pattern dictated by the colored buttons on screen. The application has many pages, including the login screen, the game screen, and the interim instruction screens (see Fig. 1)

Fig. 1. Game screen for the voice input section. Users must say the number corresponding to the square that lights up.

The application collects session data automatically, and stores that data on the server. The application can also perform automatic analysis of the collected data, and build plots as well as perform ANOVA testing, to be used in future sections.

3.2 Experimental Design

Participants were recruited on a voluntary basis to participate in the study. There were 5 volunteers. Participants were tasked with playing a game of Simon Says with each modality, voice control (VC), and mouse control (MC). Each modality was played once for each participant. Each round lasted either until 20 moves were made, or if the participant failed the pattern 2 times. Participants are measured on their score, and how many rounds they survived. Data is recorded automatically by the application for the purpose of directly comparing the MC and VC modalities.

Data analysis was conducted using an ANOVA test to compare the results between mouse control and voice control and assess the variance between the modalities. For each modality, we collected the average of each measured variable for both MC and VC. The averages were compared within subjects for each group. Analysis of these tests are shown in the “Results” section.

4 RESULTS

ANOVA Results (No Interaction):

	sum_sq	df	F	PR(>F)
Test	3.919053	1.0	3.594589	0.075098
user_id	3.567430	1.0	3.272077	0.088189
Residual	18.534498	17.0	NaN	NaN

ANOVA Results (Interaction):

	sum_sq	df	F	PR(>F)
Test	3.919053	1.0	3.575072	0.076903
user_id	3.567430	1.0	3.254312	0.090095
Test:user_id	0.995034	1.0	0.907698	0.354898
Residual	17.539464	16.0	NaN	NaN

ANOVA Results for Rounds Survived (No Interaction):

	sum_sq	df	F	PR(>F)
Test	3.200000	1.0	0.291539	0.596239
user_id	42.003883	1.0	3.826800	0.067075
Residual	186.596117	17.0	NaN	NaN

ANOVA Results for Rounds Survived (Interaction):

	sum_sq	df	F	PR(>F)
Test	3.200000	1.0	0.332289	0.572335
user_id	42.003883	1.0	4.361703	0.053096
Test:user_id	32.513592	1.0	3.376226	0.084787
Residual	154.082524	16.0	NaN	NaN

Fig. 2. ANOVA tests performed using the collected data. No Interaction shows the results just between voice and mouse control for each user, and Interaction shows the comparison between users.

4.1 Average Scores

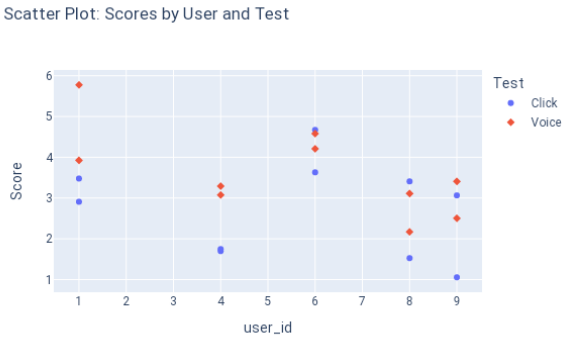

Fig. 3. A scatter plot depicting the scores obtained for each user ID.

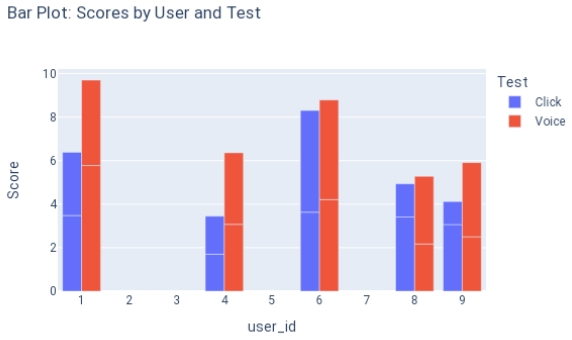

Fig. 4. A bar graph depicting the scores obtained for each user ID

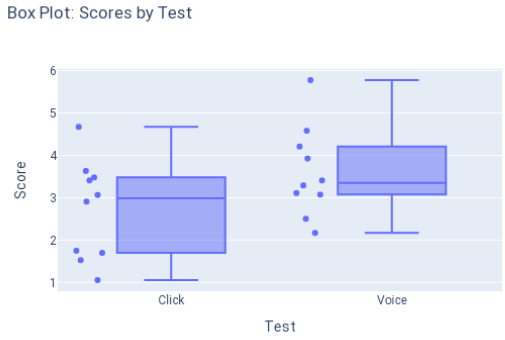

Fig. 5. Box plot depicting average scores for each modality across all users.

The first metric measure was the overall score of each test per user. Looking at Fig. 3 and Fig. 4, it is shown that participants usually performed much better using the VC modality. Each participant scored higher on the voice trial than the mouse input trial. This trend continues in the Fig. 5 box plot, showing that the top 75% of the voice scores are higher than the top 50% of the mouse input trials. From this, it can be inferred that sound-based inputs show a slight increase in performance for each participant.

For the ANOVA tests in Fig. 2, there are a couple of things to look at here. The p-value measuring the variance for each test was 0.075098, which is greater than $p = 0.05$. This means that the data showed no significant difference between the two modalities, and thus no conclusion can be made from the observed data.

Looking at the interaction ANOVA test, the observed p-value was 0.354898. This shows that there was no significant difference between the data for each participant in the study. As such, it can be stated that the results shown are not the result of any sort of confounding bias.

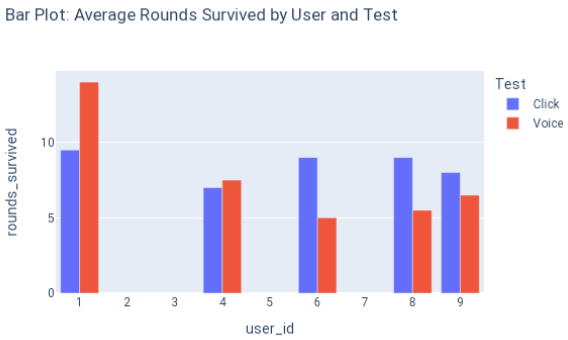

Fig. 6. A bar graph depicting how many rounds each participant lasted.

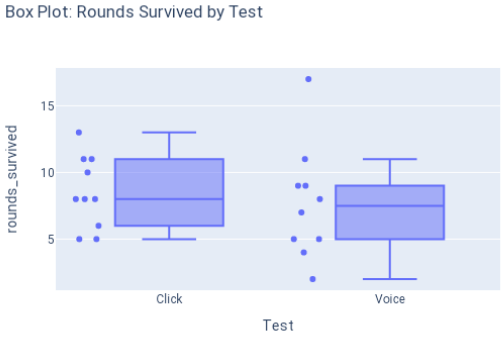

Fig. 7. Box plot depicting average rounds survived by modality.

4.2 Rounds Survived

The second variable measured was how many rounds each player survived before losing all their lives. Looking at the bar graph in Fig. 6, it seems the results on this from are quite mixed. However, by looking at the Boxplot in Fig. 7, it shows that the average round survival for mouse control is just slightly higher. What exactly is cause for this increase in survival can be all but surmised, however the general interpretation is that MC participants were less prone to losing by virtue of error.

Going back to the ANOVA tests seen in Fig. 1, the p-value depicting the variance between modalities for each user is stated to be 0.596239. This would again be greater than p = 0.05, showing that it is impossible to observe the difference in performance between the modalities is anything significant. As such, there is nothing that can be concluded from this data.

The p-value for the between-users comparison for round survival is 0.084787. This again shows the lack of confounding bias between each trial run of the experiment.

Overall, the results are largely inconclusive and show no real discernible differences between mouse control and voice control. This is congruent with some of the previous works [11]

5 DISCUSSION

The results show no significant difference between the two modalities. This lines up with the previous work that compared voice and mouse inputs [11]. The takeaway is that the two different modalities show no differences when performing a simple task that does not rely on the individual strengths of either. As for gathering a baseline difference between the two modalities, it has become abundantly clear that no differences can be observed under the circumstances of this study.

6 LIMITATIONS

This study has several limitations. First, the participant pool was entirely too small to gather any meaningful data. If a similar study is conducted in the future, a larger and more diverse population should be procured.

Second, participants running the experiment on their own set-ups, without direct supervision. For each participant, variations such as their screen resolution or their microphone quality may have impacted the results by introducing varying levels of delay or inaccuracy to their inputs. This lack of standardization allowed for more convenience given the circumstance. However, were a similar study to be conducted, these factors should be better controlled.

7 CONCLUSIONS AND FUTURE WORK

This study compared the accuracy and ease of use of two modalities at playing a simple memory game. The results showed no significant difference between the modalities. Looking into the future, it may be interesting to analyze the extent to which voice input can influence immemrsion in video games through a study of that nature. It could even perhaps implement AR and VR elements to truly make the most of the degrees of freedom provided to the user. Overall, there is much room for improvement when it comes to implementing voice control in gaming.

ACKNOWLEDGMENTS

We would like to acknowledge Doctor Francisco Ortega for their continued support.

REFERENCES

[1] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-Controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 77–81. https://doi.org/10.1145/3379172.3391728

[2] Tony Ayres and Brian Nolan. 2006. Voice activated command and control with speech recognition over WiFi. *Science of Computer Programming* 59, 1 (2006), 109–126. https://doi.org/10.1016/j.scico.2005.07.007 Special Issue on Principles and Practices of Programming in Java (PPPJ 2004).

[3] J.A. Bilmes, J. Malkin, Xiao Li, S. Harada, K. Kilanski, K. Kirchhoff, R. Wright, A. Subramanya, J.A. Landay, P. Dowden, and H. Chizeck. 2006. The Vocal Joystick. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660098

[4] Chris Carter and Robert Graham. 2000. Experimental Comparison of Manual and Voice Controls for the Operation of in-Vehicle Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44, 20 (2000), 3–286–3–289. https://doi.org/10.1177/154193120004402016 arXiv:https://doi.org/10.1177/154193120004402016

[5] Kevin Christian, Bill Kules, Ben Shneiderman, and Adel Youssef. 2000. A Comparison of Voice Controlled and Mouse Controlled Web Browsing. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies* (Arlington, Virginia, USA) *(Assets '00)*. Association for Computing Machinery, New York, NY, USA, 72–79. https://doi.org/10.1145/354324.354345

[6] Susumu Harada, Jacob O. Wobbrock, and James A. Landay. 2011. Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 11–29.

[7] Grand Research Inc. 2022. Video Game Market Size, Share & Trends Analysis Report By Device (Console, Mobile, Computer), By Type (Online, Offline), By Region (North America, Europe, Asia Pacific, Latin America, MEA), And Segment Forecasts, 2022 - 2030. , 300 pages. Retrieved April 14, 2023 from https://www.grandviewresearch.com/industry-analysis/video-game-market

[8] Chris Kikel. 2022. A Brief History of Voice Recognition Technology. Retrieved April 14, 2023 from https://www.totalvoicetech.com/a-brief-history-of-voice-recognition-technology/

[9] Sarita Sarita and Kiran Kumar Kaki. 2013. Mouse Cursor's Movements using Voice Controlled Mouse Pointer. *International Journal of Computer Applications* 71 (2013), 27–34.

[10] Nawar Shara, Margret V. Bjarnadottir, Noor Falah, Jiling Chou, Hasan S. Alqutri, Federico M. Asch, Kelley M. Anderson, Sonita S. Bennett, Alexander Kuhn, Becky Montalvo, Osirelis Sanchez, Amy Loveland, and Selma F. Mohammed. 2022. Voice activated remote monitoring technology for heart failure patients: Study design, feasibility and observations from a pilot randomized control trial. *PLOS ONE* 17, 5 (05 2022), 1–14. https://doi.org/10.1371/journal.pone.0267794

[11] Cagkan Uludagli and Cengiz Acarturk. 2018. User interaction in hands-free gaming: A comparative study of gaze-voice and touchscreen interface control. *Turkish Journal of Electrical Engineering and Computer Sciences* 26 (07 2018). https://doi.org/10.3906/elk-1710-128

[12] Jie Zhang, Ji Zhao, Shuanhu Bai, and Zhiyong Huang. 2004. Applying speech interface to Mahjong game. In *10th International Multimedia Modelling Conference, 2004. Proceedings.* 86–92. https://doi.org/10.1109/MULMM.2004.1264971

Received 22 March 2023; revised 14 April 2023