Voice-Augmented Virtual Reality Interface for Serious Games

Farhan Aslam
Department of Computer Science, University of Calgary
Calgary, Alberta, Canada, T2N1N4
farhan.aslam2@ucalgary.ca

Richard Zhao
Department of Computer Science, University of Calgary
Calgary, Alberta, Canada, T2N1N4
richard.zhao1@ucalgary.ca

***Abstract*—This paper introduces a Voice-Augmented Virtual Reality (VR) Interface designed to enhance the experience of interacting with 3D models in virtual reality environments, specifically targeting serious games and learning environments. Leveraging natural language as the primary mode of interaction, this paper presents a framework and a study that compares the voice-augmented system with a conventional hand controller version across seven fundamental 3D interaction tasks. The research explores a wide range of object interactions and offers a comprehensive voice-augmented VR interface. The system presents a more accessible and equitable alternative to controller-based interactions. This is demonstrated by a user study that compares the experiences of male and female participants engaging with the system.**

***Index Terms*—virtual reality, locomotion, accessibility, voice interaction.**

I. Introduction

Natural language interaction stands as the most instinctive and user-friendly mode of communication [1]. Given its prevalence in various daily scenarios, technologies incorporating this intuitive form of interaction possess a lower entry barrier, thus enhancing accessibility for diverse user groups. Widely adopted voice assistive technologies such as Siri and Alexa exemplify the widespread acceptance of language-based interaction. Recent advancements in natural language processing have significantly expanded the capabilities of speech interfaces, encompassing understanding spoken text [2], [3], [4], processing natural language [5], [6], [7], generating text [8], [9], and producing spoken words [10], [11], [12]. Within the realm of virtual reality (VR), speech technologies are gaining traction, being employed in conjunction with gestures for 3D scene navigation [13], [14], [15], [16], multimodal data exploration [17], [18], and as a control feature in various systems [19], [20].

Diverse approaches to VR interactions have been explored, each exhibiting distinct characteristics [21], [22]. Notably, voice-based interaction is underrepresented among these approaches, despite its intuitiveness. When improving the accessibility of an application [23] or especially in scenarios where traditional controllers or gestures are impractical, such as in sterile conditions [22], or situations where hands are occupied with secondary tasks [24], [25], users are unable to utilize controllers or gestures, suggesting alternative solutions. Among these alternatives, employing natural language emerges as arguably the most instinctive solution. With a predominant focus on the study of voice, gaze, and head movements [21], voice input serves as an interaction technique in various contexts [16] [20]. However, its implementation is often ad hoc and its effectiveness as an alternative means of VR interaction has yet to be systematically compared to conventional hand controller interactions in terms of performance and preference.

This paper introduces a voice-augmented VR Interface designed for interacting with 3D models, leveraging voice as the primary mode of interaction in a virtual environment, accessible to users with mobility impairment. Our study compares this voice-augmented system with a conventional hand controller version across seven fundamental scenarios encountered during 3D interaction (scaling, locomotion, selection, zooming, rotation, positioning, and collapsing a 3D model). Expanding on prior research, including the exploration of voice and gaze in hands-free scenarios [26], [27], our study examines a wider range of interactions. It presents a comprehensive voice-augmented VR interface that not only offers equivalent interaction but, in some instances, such as selection, provides easier alternative to hand-based interactions. Furthermore, we present a general framework for implementing voice commands that are applicable across different applications. Our interface addresses accessibility concerns, catering to individuals with mobility impairments who may face challenges utilizing traditional controls with their hands. We also address the issue of equity where women were often under-represented in VR studies [28].

In summary, the contributions of this paper are as follows:

- A framework for voice commands and the development of a voice-based tool facilitating the interactions of 3D objects in VR, accessible to users with mobility impairment.
- A comparative analysis of voice-augmented versus hand controller interactions across seven object interaction tasks in VR, showing the advantages of voice-augmented interactions.
- A comparative analysis of male and female participants in interacting with a voice-augmented VR environment, showing equitable access across the gender divide.

979-8-3503-5067-8/24/$31.00 ©2024 IEEE

II. Related Works

A. Hands-free Interactions in VR

In certain situations, such as medical environments with strict hygiene standards [22], [23], hands-free interactions become crucial. Monteiro et al. [21] offered an overview identifying key hands-free interaction techniques, the primary tasks addressed, and the metrics employed. The authors highlight voice, eye, and head as the frequently studied interaction methods. Voice-based interactions are typically categorized into two main types. First, there are systems utilizing simple one-word voice commands [29], such as 'open' or 'close.' Second, there are systems capable of recognizing and processing complete sentences [30], [31]. Notably, voice-based systems do not necessarily have to rely on speech alone; they can also respond to specific sounds [32]. For instance, Zielasko et al. [33] used a whistle sound as a start or stop command.

Eye tracking stands out as a popular method for hands-free interactions in virtual environments. By monitoring the eye position in real time, users can select or point to virtual items without moving their head [34]. Additionally, tracking eye gestures such as blinking [35] or closing the eyes [36] can confirm a selection. However, using the eyes as an interaction tool can pose challenges, as natural eye movements may be misinterpreted as undesired commands [37]. Head tracking, involving a nod or shake, is another means to confirm or deny specific interactions [38]. A dynamic indicator moving with the head position can uniquely select items in the digital world by holding the indicator over an item for a specific duration [39], [40]. Yet, in learning scenarios, head and eye tracking may be impractical, potentially distracting users from the subject of study.

Other less common alternatives for hands-free interaction encompass foot tracking [41], brain activity tracking [38], and body tracking [29]. Each approach carries its own set of considerations and implications in various contexts.

B. Speech-based Interactions in VR

In VR applications, speech interfaces manifest in diverse forms, with a prevalent utilization observed in command interfaces. Users can access a predefined set of commands articulated through natural language. Upon detecting speech, the system triggers the intended action, used in scenarios like voice-controlled positioning of virtual implants during surgical planning [20]. Voice interaction proves particularly advantageous for immersive data visualization, given its intuitive and quicker application compared to keyboard input [35], [42]. Explorations into combining speech with gestures, head movements, and gaze reveal an augmented user presence in the VR experience [31]. Dialogue interfaces find practical application in interactions with virtual avatars and agents [43], [44]. For instance, Morotti et al. [45] employed a voice assistant in a virtual reality fashion shopping experience, while Chilufya and Arvola [46] designed a virtual receptionist delivering information via speech in various rooms of a university building. Osking et al. [47] investigated the influence of speech and visual elements on narrative within a VR experience, discovering that the use of voice control heightens emotional impact on users. Voice and sound emerge as alternative channels, fostering inclusivity in virtual experiences. For instance, Ferracani et al. [48] made museums navigable for individuals with motor disabilities through voice-based interaction. Our work aligns with these systems and provides a VR learning environment solely reliant on voice interactions.

C. Accessibility in VR

This section explores dimensions of accessibility in VR. Chen et al. [49] investigated the effectiveness of different timing strategies and modalities in disambiguating gestures for VR interactions. Speech is identified as having the lowest selection errors, while head gaze exhibits higher accidental activation. The research contributes insights into designing effective disambiguation techniques and highlights scalability concerns for broader interactive workflows. Another tool, VoiceDraw [50], introduces a hands-free voice-driven drawing application tailored for individuals with motor impairments. The study showcases enhanced control, fluidity, and creative expression through continuous vocal parameter manipulation. VoiceDraw offers a new dimension of creative expression for users with motor impairments, showcasing the potential of voice-driven applications in improving accessibility. Examining the landscape of 360° video players [51], findings reveal a general lack of comprehensive accessibility support among existing video players. The study underscores the need for further research in 3D VR environments and suggests potential areas for improvement. In alignment with the discussed works above, our work represents a dedicated effort to enhance accessibility, particularly catering to individuals with limited hand mobility or unfamiliarity with hand controllers. Our system introduces a voice-augmented interface designed to facilitate the learning experience in VR for users who may face mobility impairment with traditional hand-based interactions. This voice interface aims to empower users to engage with VR educational content seamlessly using their voice, providing an improved inclusive and accessible environment.

D. VR Learning Environments and Serious Games

VR has been used in the educational context and serious games for many different fields, including manufacturing [52], biology [53], [54], chemistry [55], among others. It has also been successfully used for training skills such as perspective-taking [56], public speaking [57], scrum agile methodology [58], for enhancing presence and motivation [59], and for drug addiction prevention [60]. In the realm of VR for serious purposes, multi-dimensional approaches have been harnessed to enrich the learning experience. One of the contributions introduces an innovative overview visualization inspired by library shelves [53]. This comprehensive visualization encompasses the liver surface, blood vessels, gall bladder, and various tumors or cysts. The interaction techniques employed include the Virtual Hand method, enabling users to grab, translate, and rotate 3D models. Bimanual interaction is

incorporated for scaling purposes, offering a dynamic and immersive learning environment. The study explores the profound implications of multi-user liver anatomy education in both virtual and augmented reality environments. Another noteworthy contribution explores the creation of an interactive 3D torso anatomy education environment [54]. Demonstrating deformable anatomy models of male and female torso organs, it leverages mid-air gesture interaction. This paper enriches the educational landscape with immersive and interactive anatomy learning experiences. A separate study explores application of immersive technology to facilitate learning the creation of molecules of chemical compounds [55]. Utilizing the VR Nanome, practical activities immerse students in a virtual organic chemistry laboratory. In a different domain, a study focuses on the comparison of voice, gesture, and controller interface methods [61]. Conducted using HoloLens, the research emphasizes the suitability of different interaction methods for future VR applications for children, with a specific focus on the voice interaction task completed with the "select" keyword. In comparison to these contributions, our proposed system extends the immersive learning experience to a broader spectrum of subjects beyond anatomy. Users can interact with any imported 3D model as part of their study material, leveraging voice commands instead of traditional hand-based interactions. Unlike the ad hoc implementation of prior work with limited voice commands, our system provides a versatile set of voice commands under a general framework that can be easily extended. This approach aligns with the broader vision of creating inclusive, accessible, and efficient VR learning environments for a diverse range of user needs.

III. Framework and Implementation

This section provides an overview of our voice-augmented framework, our voice VR interface implemented in a VR environment, and lists the tools and hardware used to implement the application.

A. Voice-Augmented Framework

We design a voice-augmented framework for supporting VR interactions, specially interactions with virtual objects. Manipulating 3D objects in VR is common for VR-based learning environments and serious games, therefore this research specifically examine this aspect of VR interactions.

The general framework of understanding of voice commands in our interface is based on recognition of intent, entity and trait [62]. In the context of VR environments, we adapt and define the following:

- **Intent** is the action or verb of the voice command
- **Entity** is the object in VR upon which the action is performed
- **Trait** is the property or amount of action based on the specific Intent

A voice command would take the form:

<Intent><optional Entity><optional Trait>

For example, in the voice command "select 34", the "select" word is the intent and "34" is the entity labelled with ID 34. As a consequence of this command entity with ID 34 would be selected.

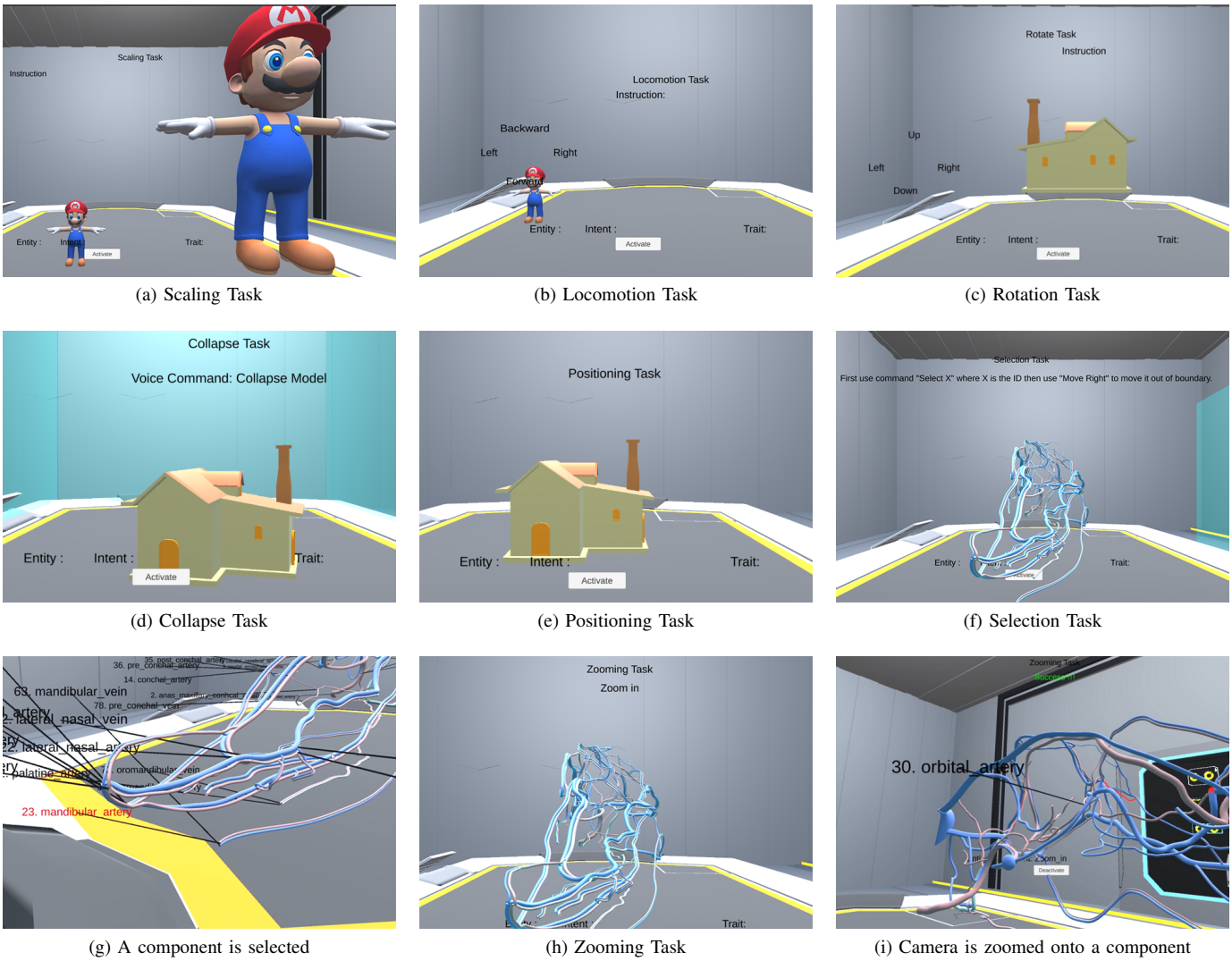
This framework is state-based. An entity selected would be memorized as the current state, so that later commands can omit the Entity and refer to the entity in the current state as the default. If the next voice command is "rotate right at 45 degrees", then entity 34 would be acted upon as the default entity, and it would be rotated to the right, as "Rotate Right" is the intent and "45" would be the trait or amount of rotation which is an angle.

In our implementation, the current state is comprised as one entity. However, this general framework does not preclude an expanded current state that includes multiple entities. The user would either need to specify a particular entity or have the command apply to all entities in the current state.

B. Voice Interface Implementation

Our VR interface is implemented in the Unity engine [63]. The environment provides users with a hands-free approach to explore and learn complex 3D models based on the aforementioned framework. The application environment resembles a typical laboratory having a table at the center on which the imported models are examined, a Unity canvas remains in front of the user to convey recognized voice commands. The system supports a comprehensive set of voice commands that can be easily extended following the framework.

These commands facilitate diverse functionalities, allowing users to interact seamlessly with the 3D model. For instance, the "annotate" command creates labeled annotations for individual components of a model (using names given in Unity), establishing a visual connection between the label and the corresponding component. "Hide annotations" clears all created annotations, while "increase/decrease plane distance" controls the visibility of annotations within a specified boundary, indicated by a semi-transparent plane. The command "collapse/merge model" alters a model's appearance, providing either a spaced-out or original view. "Pronounce names" directs the camera to each component of the model on the screen, highlighting and audibly pronouncing their names. Users can stop the pronunciation with the "stop pronunciation" command. To interact with specific components, "select" enables users to choose individual components of the model. Navigational commands like "move forward/backward/left/right" facilitate movement of model or selected component by a distance specified by Trait, and "zoom in/out" brings the user closer or further. Rotation commands ("rotate left/right/up/down") rotates the object by a specified angle in the given direction. Users can scale components with "scale up/down" and move them up or down with "up/down". Cross-sectional views are accessible through "cross section backward/forward" and "cross section up/down," showing vertical and horizontal cross sections, respectively. "Return to home position" restores the model and selected component to their original state. Additional functionalities include "copy," "unlock/lock view" to freeze or unfreeze the view, and "reset



(a) Scaling Task (b) Locomotion Task (c) Rotation Task

(d) Collapse Task (e) Positioning Task (f) Selection Task

(g) A component is selected (h) Zooming Task (i) Camera is zoomed onto a component

Fig. 1. Our user study involves seven tasks: (a) scaling a model to match a target, (b) moving a model to a target location, (c) rotating a model to a target orientation, (d) breaking a model into its individual components, (e) restoring displaced components to their original locations, (f) selecting a specific component, and (h) zooming to a specific component. (g) shows the result of having a component of a model selected and its label highlighted in red. (i) shows the result of having zoomed onto a component. Several 3D models were used, including a Mario character, a house with components, and the veins of an alligator head. The models were chosen to suit the increasing complexity of the tasks, and to provide some variety to keep user engagement.

environment" to restore the model to its original imported state.

C. Software Setup

To implement our VR interface, we needed a development environment capable of building VR applications. Our choice was the Unity engine, offering broad compatibility, cross-platform development, and strong community support. Selecting the right speech recognition application programming interface (API) involved considering factors such as licensing and Unity support. We opted for the Wit.AI API, meeting our criteria of being freely available and easily integrated with Unity. In the main scene of the application, scripts were written to implement functionalities for each supported voice command. To enhance recognition, Wit.AI provided the ability to fine-tune its model with custom voices. We fine-tuned it with twenty-one freely available voices from IBM Watson[1] and ReadLoud[2]. covering a wide range of accents featuring male and female voices. Our application easily supports the addition of new voice commands following our framework, and the use of OpenXR makes it cross-platform.

IV. User Study

Interface usability evaluation often relies on metrics such as satisfaction, efficiency, and efficacy, commonly gathered through tailored questionnaires to collect user feedback on preferences and interface usage. Some studies employ validated questionnaires for this purpose [21], [64]. Similarly, we used pre and post study questionnaires to support our analysis.

[1]https://www.ibm.com/products/text-to-speech
[2]https://readloud.net/

A. Study Setup

The user study was conducted in a controlled in-person lab environment, with identical equipment provided to each participant to ensure comparable conditions. To maintain a focused environment for processing participants' voices, only the participant was allowed to speak during the study and additional signs were placed outside and, in the lab, to avoid interruptions. Participants were informed that questions during the study were limited to special cases preventing unintended command triggers. Any uncertainties or questions were addressed before initiating the actual study. After a brief introduction to the first version of the study and equipment, participants were immersed in a virtual world. Following the first part of the study participants were given a short break and briefing for the second version. Post study, participants filled out a questionnaire regarding discomfort, comfort, presence, and usability. The entire process, including preparations, the study, and the post-study questionnaire, took approximately 60 minutes per participant.

We utilized a laptop (11th Gen Intel(R) Core (TM) i7-11800H @ 2.30GHz) and a desktop PC featuring a 12th Gen Intel(R) Core (TM) i7-12700H @ 2.10GHz with Nvidia GeForce GTX 1060 graphics. Oculus Rift S was used for this study and was connected to the desktop PC. The laptop was used for survey collection.

B. Procedure

We implemented two versions of the same virtual environment to compare hand and voice interactions: Voice Version (VV) and Hand Version (HV). Each version comprised of seven tasks. In each task, a participant needed to complete five objectives. In both versions, a practice task preceded each main task. The addition of a practice task ensured that all participants felt comfortable and confident with both versions, and no time limits were imposed on their practice. This approach allowed participants to fully understand the tasks, regardless of their prior familiarity with VR interactions. Participants were asked to complete both versions. To mitigate the learning effect in doing the second version of the study, half of the participants were randomly assigned to begin with VV, and the other half begin with HV.

To compare the two modes of interactions, seven tasks covering common scenarios in interactions with 3D models were identified. These tasks included: 1. Scaling (increasing or decreasing size of the object); 2. Locomotion (moving the object to target locations); 3. Rotation (rotating the object to specific angles); 4. Collapse (breaking the object into individual components); 5. Positioning (restoring the original position of displaced components of an object); 6. Selection (selecting an individual component in a complex model); and 7. Zooming (adjusting the view to have a closer look at a specific component) (Figure 1). VV utilized the framework as described above, while HV deployed traditional hand-controller-based interactions, with objects selected by pointing the controller at them, and manipulated by pressing and holding different buttons on the controller.

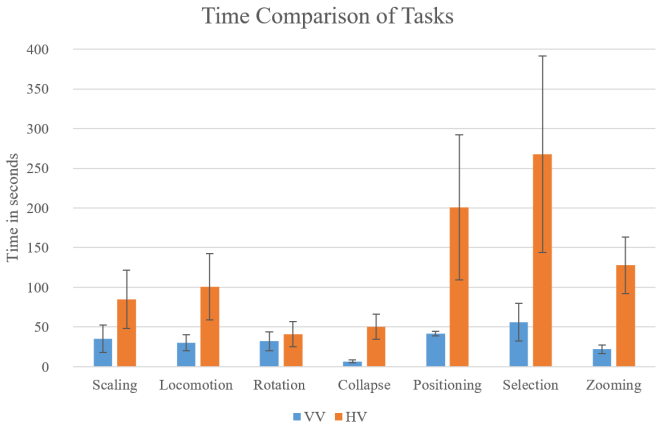


Fig. 2. Completion time comparison between VV and HV over the seven tasks, averaged over all participants. Error bars represent one standard deviation. Numeric values are listed in Table III.

V. Results and Discussions

Each participant was assigned a unique ID to ensure anonymity of the study data. In total, 26 participants were recruited from a university. There were 12 self-identified male and 14 self-identified female participants, aged 18 to 29. Participants reported White, Latino, South Asian, East Asian, and Middle Eastern as their ethnic backgrounds. All participants spoke English at native, advanced, or intermediate levels. Only two participants responded to sensitivity to VR sickness as "somewhat" while the rest responded "no."

After the study, participants completed the post study questionnaire. Twenty-five participants agreed that interacting with 3D models was easier using voice commands. Furthermore, all participants were of the view that it took less time to complete the tasks in VV. Regarding user discomfort/fatigue/dizziness, participants provided ratings on a scale of 1 to 10, where 10 signified maximum discomfort/fatigue/dizziness and 1 minimum. Notably, all participants indicated ratings of 1 or 2 for VV while the responses for HV exhibited a more varied distribution between 3 and 9. Five participants indicated a level of 3, another five participants at 4, eight participants at 5, four participants at 6, two participants at 7, one participant at 8, and one participant at 9. This suggests a broader and higher level of discomfort/fatigue/dizziness experienced by participants using HV. Finally, nearly all participants (25 out of 26) expressed that they felt less confused when completing tasks with voice commands compared to using hand controllers. This indicated a higher level of clarity and ease of understanding in utilizing voice-augmented interactions.

A. Quantitative Analysis

The common measure across all tasks is time taken to complete each task. The comparative analysis between the VV and HV reveals notable advantages of voice-augmented interactions across various dimensions. Figure 2 shows the comparison of the completion times of each of the seven tasks, averaged over the 26 participants. VV outperformed HV in every task, with the results statistically significant at the 95%

confidence level using two-tailed paired t-tests. We observe that VV excelled in tasks requiring selecting components of a complex model and zooming to such a component. For example, the average time taken for the selection task in VV was 56.3 seconds, significantly lower than HV's 267.6 seconds.

Precision of the submitted results is measured in terms of the total errors made (Table I). Since users were allowed to submit a task result without precisely meeting the requirements in scaling, locomotion, positioning, and zooming, we measured precision of the submitted results. For the scaling task, differences between the target and user-submitted scales were recorded. Position differences between user-submitted and target positions were logged for locomotion, positioning, and zooming tasks. The ability of VV to specify precise commands was an advantage HV cannot match.

Another measure is the total number of mistakes made by participants due to taking an action that did not lead to the goal (Table II): the number of wrong selections was recorded for the selection task; for locomotion, positioning, and zooming tasks, movement mistakes were counted when users moved in directions other than the correct one; rotation mistakes were considered for rotation and zooming tasks when users rotated in the wrong direction (every wrong input counted as one). In VV, mistakes were recorded for all tasks whenever users gave a voice command that was not leading towards completing the tasks. The results indicated finer control and accuracy achieved through voice commands. However, VV had a caveat of occasionally recognizing the wrong voice command (when the participant said one command and the voice recognition system understood it as another), albeit still at a much lower frequency compared to the mistakes made in HV. The total numbers of wrong voice recognition of all 26 participants were 56 (scaling), 12 (locomotion), 20 (rotation), 0 (collapse), 4 (positioning), 16 (selection), and 5 (zooming). We believe that this could be improved with better voice recognition algorithms, and fine-tuning with a participant's voice instead of generic voices.

TABLE I
TOTAL ERRORS MADE BY ALL PARTICIPANTS DURING THE ENTIRE STUDY, MEASURED IN UNITY IN-GAME UNITS.

Errors	Voice Version	Hand Version
Scaling Task Size Errors	0.00	0.29
Locomotion Task Position Errors	0.00	22.12
Positioning Task Position Errors	0.00	3.00
Zooming Task Position Errors	0.00	10.54

Lastly, we present the results of a comparative analysis of female and male participants in interacting with a voice-augmented VR environment (Table III). Researchers have consistently pointed out the under-representation of female participants in VR research [28]. This could have substantial adverse effects on diversity and inclusion in VR-based serious games and learning environments. Past research has shown that men and women might interact with VR in different ways [65]

TABLE II
TOTAL MISTAKES MADE BY ALL PARTICIPANTS DURING THE ENTIRE STUDY.

Mistakes	Voice Version	Hand Version
Locomotion Task Movement	5	78
Rotation Task Rotation	4	2390
Positioning Task Movement	0	121
Selection Task	0	60
Zooming Task Rotation	0	1509
Zooming Task Movement	0	171

TABLE III
TIME COMPARISON BY GENDER, AVERAGED ACROSS PARTICIPANTS. TIME MEASURED IN SECONDS. STANDARD DEVIATION IN BRACKETS.

Tasks	All Participants		Male Participants		Female Participants	
	VV	HV	VV	HV	VV	HV
Scaling	35.2 (17.4)	84.7 (36.8)	34.5 (19.5)	83.5 (46.6)	35.9 (15.4)	85.8 (25.5)
Locomotion	30.4 (9.9)	100.7 (41.4)	31.0 (11.1)	83.0 (30.8)	29.8 (8.7)	115.9 (43.3)
Rotation	31.9 (11.7)	40.6 (15.9)	32.5 * (10.7)	38.0 * (17.3)	31.4 (12.4)	42.9 (14.2)
Collapse	6.6 (1.7)	50.4 (15.7)	6.8 (1.8)	46.0 (15.8)	6.5 (1.6)	54.3 (14.6)
Positioning	41.4 (2.9)	200.6 (91.4)	40.5 (2.5)	164.8 (59.0)	42.1 (3.0)	231.2 (102.4)
Selection	56.3 (23.8)	267.6 (123.7)	53.4 (28.3)	239.3 (112.2)	58.7 (18.8)	291.9 (127.8)
Zooming	21.8 (5.1)	127.8 (35.8)	22.6 (6.0)	114.2 (32.9)	21.2 (4.2)	139.5 (34.0)

and that women are more susceptible to simulation sickness than men [66]. This could be the result of VR devices not designed to fit the facial features of women [67]. Therefore, it is important for us to analyze any differences between female and male participants when proposing a new framework in VR. Table III shows that VV consistently outperformed HV in every task, across the gender divide, with statistical significance shown in every case, except for male participants in the rotation task (marked in table with *). Notably, even though in HV, male participants took less time on average than female participants for all tasks, this is not the case for VV. For VV, there are no statistical differences between female and male participants for any task. This finding indicates that a voice-augmented VR system is equally accessible and user-friendly for both female and male users, thereby reducing entry barriers and enhancing inclusivity for women using VR environments. Examining the standard deviations across tasks provides insights into the consistency of performance. VV consistently exhibited lower standard deviations across tasks, indicating a more reliable performance compared to HV.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a voice-augmented framework and implementation for VR-based interactions in virtual environments. In the light of the results discussed, the voice-augmented interface not only demonstrated low error rate but also exhibited superior precision and efficiency, supporting our contribution to provide hand controller equivalent interactions

in VR. We also showed the female users are not disadvantaged by the voice-augmented system compared to male users. These findings underscore the potential of how voice-augmented interactions can be used to provide an accessible VR interfaces for all users, particularly in educational contexts where precise and intuitive interactions with 3D models are crucial for a seamless learning experience.

While our study has provided valuable insights into the comparative analysis of speech-based and hand controller interactions in VR environments, future work could focus on refining voice recognition algorithms to improve the accuracy of the voice interface. Investigating machine learning techniques may contribute to more robust and nuanced voice command understanding. Further efforts can be directed towards making VR interfaces more accessible for a wider range of user groups, including those with varying levels of physical disabilities, by incorporating eye-tracking technologies. Adapting the interface to accommodate users with different needs and preferences would contribute to a more inclusive learning environment. Extending the application of the VR voice interface to cover diverse domains beyond object interactions could open new possibilities and contribute to the innovation of accessible VR interfaces for serious games and virtual learning environments.

ACKNOWLEDGMENT

This research was supported by the NSERC Discovery Grant. We thank members of the Serious Games Research Group and the anonymous reviewers.

REFERENCES

[1] D. Perez-Marin and I. Pascual-Nieto, *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global, 2011.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6465–6469.

[4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *Advances in neural information processing systems*, vol. 32, 2019.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[10] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[13] M. Baxter, A. Bleakley, J. Edwards, L. Clark, B. R. Cowan, and J. R. Williamson, ""you, move there!": Investigating the impact of feedback on voice control in virtual environments," in *Proceedings of the 3rd Conference on Conversational User Interfaces*, 2021, pp. 1–9.

[14] M. Friedrich, S. Langer, and F. Frey, "Combining gesture and voice control for mid-air manipulation of cad models in vr environments," *arXiv preprint arXiv:2011.09138*, 2020.

[15] A. Grinshpoon, S. Sadri, G. J. Loeb, C. Elvezio, and S. K. Feiner, "Hands-free interaction for augmented reality in vascular interventions," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018, pp. 751–752.

[16] J. Sin and C. Munteanu, "Let's go there: Voice and pointing together in vr," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–3.

[17] B. Lee, A. Srinivasan, J. Stasko, M. Tory, and V. Setlur, "Multimodal interaction for data visualization," in *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, 2018, pp. 1–3.

[18] J. Wang, J. Li, and X. Shi, "Integrated design system of voice-visual vr based on multi-dimensional information analysis," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 1–8, 2021.

[19] C. Li and B. Tang, "Research on voice interaction technology in vr environment," in *2019 International Conference on Electronic Engineering and Informatics (EEI)*. IEEE, 2019, pp. 213–216.

[20] H.-R. Rantamaa, J. Kangas, M. Jordan, H. Mehtonen, J. Mäkelä, K. Ronkainen, M. Turunen, O. Sundqvist, I. Syrjä, J. Järnstedt *et al.*, "Evaluation of voice commands for mode change in virtual reality implant planning procedure," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 11, pp. 1981–1989, 2022.

[21] P. Monteiro, G. Gonçalves, H. Coelho, M. Melo, and M. Bessa, "Hands-free interaction in immersive virtual reality: A systematic review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2702–2713, 2021.

[22] A. Mewes, B. Hensen, F. Wacker, and C. Hansen, "Touchless interaction with software in interventional radiology and surgery: a systematic literature review," *International journal of computer assisted radiology and surgery*, vol. 12, pp. 291–305, 2017.

[23] Y. Li, "Living in a virtual world: how vr helps disabled people to explore the world," in *International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2021)*, vol. 12153. SPIE, 2021, pp. 210–216.

[24] J. Austerjost, M. Porr, N. Riedel, D. Geier, T. Becker, T. Scheper, D. Marquard, P. Lindner, and S. Beutel, "Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments," *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 23, no. 5, pp. 476–482, 2018.

[25] D. W. Carruth, C. R. Hudson, C. L. Bethel, M. Pleva, S. Ondas, and J. Juhar, "Using hmd for immersive training of voice-based operation of small unmanned ground vehicles," in *International conference on human-computer interaction*. Springer, 2019, pp. 34–46.

[26] D. Calandra, F. G. Pratticò, and F. Lamberti, "Comparison of hands-free speech-based navigation techniques for virtual reality training," in *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2022, pp. 85–90.

[27] J. Hombeck, H. Voigt, T. Heggemann, R. R. Datta, and K. Lawonn, "Tell me where to go: Voice-controlled hands-free locomotion for virtual reality systems," in *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2023, pp. 123–134.

[28] T. C. Peck, L. E. Sockol, and S. M. Hancock, "Mind the gap: The underrepresentation of female participants and authors in virtual reality research," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 5, pp. 1945–1954, 2020.

[29] B. L. Schroeder, S. K. Bailey, C. I. Johnson, and E. Gonzalez-Holland, "Presence and usability do not directly predict procedural recall in

virtual reality training," in *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part II 19*. Springer, 2017, pp. 54–61.

[30] L. Alfaro, R. Linares, and J. Herrera, "Scientific articles exploration system model based in immersive virtual reality and natural language processing techniques," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 7, 2018.

[31] A. Marcus and W. Wang, *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II*. Springer, 2019, vol. 11584.

[32] M. Sra, X. Xu, and P. Maes, "Breathvr: Leveraging breathing as a directly controlled interface for virtual reality games," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[33] D. Zielasko, N. Neha, B. Weyers, and T. W. Kuhlen, "A reliable non-verbal vocal input metaphor for clicking," in *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2017, pp. 40–49.

[34] J. Blattgerste, P. Renner, and T. Pfeiffer, "Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views," in *Proceedings of the Workshop on Communication by Gaze Interaction*, 2018, pp. 1–9.

[35] D. Kumar and A. Sharma, "Electrooculogram-based virtual reality game control using blink detection and gaze calibration," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2016, pp. 2358–2362.

[36] J. Kim, J. Cha, H. Lee, and S. Kim, "Hand-free natural user interface for vr hmd with ir based facial gesture tracking sensor," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, 2017, pp. 1–2.

[37] R. J. Jacob, "The use of eye movements in human-computer interaction techniques: what you look at is what you get," *ACM Transactions on Information Systems (TOIS)*, vol. 9, no. 2, pp. 152–169, 1991.

[38] C. Pavlatos and V. Vita, "Linguistic representation of power system signals," *Electricity Distribution: Intelligent Solutions for Electricity Transmission and Distribution Networks*, pp. 285–295, 2016.

[39] Y. Yan, Y. Shi, C. Yu, and Y. Shi, "Headcross: Exploring head-based crossing selection on head-mounted displays," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 1, pp. 1–22, 2020.

[40] Y. Yan, C. Yu, X. Yi, and Y. Shi, "Headgesture: Hands-free input approach leveraging head movements for hmd devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–23, 2018.

[41] K. Minakata, J. P. Hansen, I. S. MacKenzie, P. Bækgaard, and V. Rajanna, "Pointing by gaze, head, and foot in a head-mounted display," in *Proceedings of the 11th ACM symposium on eye tracking research & applications*, 2019, pp. 1–9.

[42] M. Lange, J. Hjalmarsson, M. Cooper, A. Ynnerman, and V. Duong, "3d visualization and 3d and voice interaction in air traffic management," in *Proceedings of the Annual SIGRAD Conference, special theme Real Time Simulations*. Citeseer, 2003, pp. 17–22.

[43] G. Ali, H.-Q. Le, J. Kim, S.-W. Hwang, and J.-I. Hwang, "Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable mixed reality environment," in *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*, 2019, pp. 47–52.

[44] S. Uchino, N. Abe, K. Tanaka, T. Yagi, H. Taki, and S. He, "Vr interaction in real-time between avatar with voice and gesture recognition system," in *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, vol. 2. IEEE, 2007, pp. 959–964.

[45] E. Morotti, L. Donatiello, and G. Marfia, "Fostering fashion retail experiences through virtual reality and voice assistants," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 338–342.

[46] E. M. Chilufya and M. Arvola, "Conceptual designing of a virtual receptionist: Remote desktop walkthrough and bodystorming in vr," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 112–120.

[47] H. Osking and J. A. Doucette, "Enhancing emotional effectiveness of virtual-reality experiences with voice control interfaces," in *Immersive Learning Research Network: 5th International Conference, iLRN 2019, London, UK, June 23–27, 2019, Proceedings 5*. Springer, 2019, pp. 199–209.

[48] A. Ferracani, M. Faustino, G. X. Giannini, L. Landucci, and A. Del Bimbo, "Natural experiences in museums through virtual reality and voice commands," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1233–1234.

[49] D. L. Chen, R. Balakrishnan, and T. Grossman, "Disambiguation techniques for freehand object manipulations in virtual reality," in *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 2020, pp. 285–292.

[50] S. Harada, J. O. Wobbrock, and J. A. Landay, "Voicedraw: a hands-free voice-driven drawing application for people with motor impairments," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, 2007, pp. 27–34.

[51] C. J. Hughes and M. Montagud, "Accessibility in 360 video players," *Multimedia tools and applications*, vol. 80, no. 20, pp. 30 993–31 020, 2021.

[52] F. Aqlan, R. Zhao, H. C. Lum, and L. J. Elliott, "Integrating simulation games and virtual reality to teach manufacturing systems concepts," in *2019 ASEE Annual Conference & Exposition*, 2019.

[53] D. Schott, P. Saalfeld, G. Schmidt, F. Joeres, C. Boedecker, F. Huettl, H. Lang, T. Huber, B. Preim, and C. Hansen, "A vr/ar environment for multi-user liver anatomy education," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2021, pp. 296–305.

[54] N. Zhang, H. Wang, T. Huang, X. Zhang, and H. Liao, "Deformable torso anatomy education with three-dimensional autostereoscopic visualization and free-hand interaction," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2022, pp. 552–553.

[55] J. Á. Ramírez and A. M. V. Bueno, "Learning organic chemistry with virtual reality," in *2020 IEEE International Conference on Engineering Veracruz (ICEV)*. IEEE, 2020, pp. 1–4.

[56] R. Band, M. Lips, J. Prawira, J. van Schagen, S. Tulling, Y. Zhang, A. A. Benaiss, I. J. van der Ham, M. Bueno, and R. Bidarra, "Training and assessing perspective taking through a hole new perspective," in *2022 IEEE Conference on Games (CoG)*. IEEE, 2022, pp. 268–275.

[57] R. Dillon, A. N. Teoh *et al.*, "Real-time stress detection model and voice analysis: An integrated vr-based game for training public speaking skills," in *2021 IEEE Conference on Games (CoG)*. IEEE, 2021, pp. 1–4.

[58] D. López-Fernández, J. Mayor, J. Pérez, and A. Gordillo, "Learning and motivational impact of using a virtual reality serious video game to learn scrum," *IEEE Transactions on Games*, 2022.

[59] F. Born, L. Graf, and M. Masuch, "Exergaming: The impact of virtual reality on cognitive performance and player experience," in *2021 IEEE Conference on Games (CoG)*. IEEE, 2021, pp. 1–8.

[60] A. Vetter, J. Büttner, and S. von Mammen, "Baked burger bash: A serious virtual reality game informing about the effects of acute cannabinoid intoxication," in *2023 IEEE Conference on Games (CoG)*. IEEE, 2023, pp. 1–4.

[61] B. Munsinger and J. Quarles, "Augmented reality for children in a confirmation task: Time, fatigue, and usability," in *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, 2019, pp. 1–5.

[62] M. Mitrevski and M. Mitrevski, "Getting started with wit.ai," *Developing Conversational Interfaces for iOS: Add Responsive Voice Control to Your Apps*, pp. 143–164, 2018.

[63] J. K. Haas, "A history of the unity game engine," *Diss. Worcester Polytechnic Institute*, vol. 483, no. 2014, p. 484, 2014.

[64] E. S. Martinez, A. S. Wu, and R. P. McMahan, "Research trends in virtual reality locomotion techniques," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2022, pp. 270–280.

[65] M. Bonfert, R. Porzel, and R. Malaka, "Get a grip! introducing variable grip for controller-based vr systems," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 604–612.

[66] K. M. Stanney, K. S. Hale, I. Nahmens, and R. S. Kennedy, "What to expect from immersive virtual environment exposure: Influences of gender, body mass index, and past experience," *Human factors*, vol. 45, no. 3, pp. 504–520, 2003.

[67] K. Stanney, C. Fidopiastis, and L. Foster, "Virtual reality is sexist: but it does not have to be," *Frontiers in Robotics and AI*, vol. 7, p. 4, 2020.