Tell Me Where To Go: Voice-Controlled Hands-Free Locomotion for Virtual Reality Systems

Jan Hombeck* University of Jena
Henrik Voigt University of Jena
Timo Heggemann University Hospital Cologne
Rabi R. Datta † University Hospital Cologne
Kai Lawonn † University of Jena

ABSTRACT

As locomotion is an important factor in improving Virtual Reality (VR) immersion and usability, research in this area has been and continues to be a crucial aspect for the success of VR applications. In recent years, a variety of techniques have been developed and evaluated, ranging from abstract control, vehicle, and teleportation techniques to more realistic techniques such as motion, gestures, and gaze. However, when it comes to hands-free scenarios, for example to increase the overall accessibility of an application or in medical scenarios under sterile conditions, most of the announced techniques cannot be applied. This is where the use of speech as an intuitive means of navigation comes in handy. As systems become more capable of understanding and producing speech, voice interfaces become a valuable alternative for input on all types of devices. This takes the quality of hands-free interaction to a new level. However, intuitive user-assisted speech interaction is difficult to realize due to semantic ambiguities in natural language utterances as well as the high real-time requirements of these systems. In this paper, we investigate steering-based locomotion and selection-based locomotion using three speech-based, hands-free methods and compare them with leaning as an established alternative. Our results show that landmark-based locomotion is a convenient, fast, and intuitive way to move between locations in a VR scene. Furthermore, we show that in scenarios where landmarks are not available, number grid-based navigation is a successful solution. Based on this, we conclude that speech is a suitable alternative in hands-free scenarios, and exciting ideas are emerging for future work focused on developing hands-free ad hoc navigation systems for scenes where landmarks do not exist or are difficult to articulate or recognize.

Index Terms: Human-centered computing—Human computer interaction (HCI)—;——Computing methodologies—Artificial intelligence—Natural language processing Speech recognition

1 INTRODUCTION

For humans, natural language is the most intuitive interaction [74]. People unconsciously use language in almost all situations of daily life to communicate goals, exchange ideas, or express themselves. Computer systems that can effortlessly understand and handle speech offer a low interaction barrier and are accessible to a large number of users out of the box. Advances in natural language processing (NLP) technologies in recent years mostly based on deep learning [71] have increased the capabilities of speech interfaces to understand spoken text [6, 84, 88], process natural language [25, 80, 105], generate text [26, 52, 78] and generate spoken words [45, 91, 101]. In virtual reality (VR), the use of speech technologies is becoming increasingly popular, for example in combination with gestures for navigation in 3D scenes [7,32,35,93], as well as for multimodal data exploration [50, 51, 99] and in systems that use speech as a control feature [53, 82].

*e-mail: jan.hombeck@uni-jena.de
†Contributed equally

With regard to virtual reality locomotion techniques, a wide range of different approaches have been studied, each with different characteristics [57]. Di Luca et al. [57] find that especially controller, vehicle and teleportation-based approaches cause users to feel more simulator sickness and that the locomotion itself feels less natural. In contrast, motion, gesture, and gaze-based approaches cause less nausea and are more similar to real-world movements. Remarkably, voice-based locomotion is *not* represented among the listed approaches, although the collection of techniques [57] is very comprehensive. In certain scenarios, such as in surgical procedures under sterile conditions [23, 64], in situations where the hands are used for a secondary task [4, 16] or when improving the accessibility of an application [31, 55], users cannot use controllers or gestures and therefore need alternatives. Of these, using natural language is perhaps the most intuitive solution. Such scenarios, referred to as *hands-free*, focus on alternative interaction techniques, with voice, gaze, and head movements being the most commonly studied [68]. Although voice input is used as an interaction technique in these contexts [82, 93], its value as an alternative method of locomotion for hands-free scenarios in sequential navigation tasks has not been compared to other hands-free methods in terms of its performance and satisfaction, efficiency, and efficacy.

In this paper, we focus on three different variants of voice as a locomotion technique that is suitable for hands-free scenarios. As such, (i) voice-based steering as a representative of steering-based locomotion, and (ii) landmark teleportation, and (iii) number teleportation as representatives of selection-based locomotion are examined in detail. We compare the techniques with the established hands-free locomotion technique of *leaning* [12, 38, 46]. The leaning technique, which closely resembles real movement, has been shown to provide a high degree of self-motion experience [38]. It can be performed with or without additional controls and especially in hands-free scenarios. Leaning usually causes a relative change in position in the direction the user is leaning and is therefore intuitively usable for sequential navigation of targets in a 3D scene. In prior work, Calandra et al. [15] studied the combination of voice and gaze in hands-free scenarios in a point-to-point navigation task. Our work extends these investigations to sequential locomotion tasks that move across multiple points toward a target as performed in several use cases. One such application for voice-controlled hands-free locomotion is in the field of medicine, for example in virtual endoscopy, as described in [43, 56, 111]. The technology enables examination of the inside of a patient's body using voice commands to navigate the virtual environment. The physician's hands can be devoted entirely to controlling medical instruments, while voice control handles moving between locations in space. Another application of voice-controlled hands-free locomotion is inspection and training in architecture and construction, as mentioned in [18, 98, 100]. Voice commands are used to quickly move around large buildings or construction sites, ask contextual questions, annotate discovered defects in construction, or trigger predefined functions during inspection via voice command. Further use cases include digital tours and museums, as described in [31, 59, 109]. Voice instructions are used to navigate and interact with exhibits. Here, voice control not only provides a more immersive, interactive, and engaging experience but also improves accessibility for individuals with mobility impairments

2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR) | 979-8-3503-4815-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/VR55154.2023.00028

2642-5254/23/$31.00 ©2023 IEEE
DOI 10.1109/VR55154.2023.00028

who are unable to physically visit these attractions or for whom the use of controls is not possible. An additional aspect we are investigating is that movements are performed in environments that may have no or non-articulable landmarks, as is the case in specific medical visualizations like blood vessels [29] or abstract molecule visualizations [95]. Our results provide important insights into the usability of speech in sequential, hands-free locomotion tasks, as well as how to deal with existing or non-existing landmarks. This provides ideas for further research on the development of interaction interfaces in hands-free scenarios.

Our contributions can therefore be summarized as follows:

- Comparison of three speech-based, hands-free locomotion techniques against leaning as an established alternative for use in steering-based locomotion or selection-based locomotion in sequential navigation tasks.
- Comparison of scenarios with and without articulable visual landmarks.

In the following Section 2, we give an overview of recent work on locomotion techniques in VR, followed by specific studies on hands-free interaction techniques as well as related work in natural language processing. Section 3 presents the methods studied, starting with the use of speech in steering-based locomotion, moving to selection-based methods, and ending with leaning. Sections 4 and 5 evaluate the user study and discuss our results considering performance measures for each technique as well as an in-depth evaluation of their values in terms of satisfaction, efficiency, and effectiveness. Finally, in Section 6, we provide an outlook on future topics that can be derived from our work, specifically addressing voice-based locomotion using ad hoc number grids as an interesting direction for future research in environments with sparse or difficult-to-articulate visual landmarks.

2 Related Work

In this section, we refer to related work that addresses locomotion techniques in VR (2.1) as well as hands-free interaction in VR (2.2) as a specific subset thereof. In addition, we refer to existing work that uses speech in VR interaction (2.3), as well as work in the NLP field that involves the recognition and synthesis of speech and text and is used in VR systems for underlying speech processing (2.4).

2.1 Locomotion in VR

Navigation in the virtual world is one of the most important aspects to give users the feeling of immersion. While the most natural way of physical movement is supported in almost every VR application, restriction within the physical space makes navigating very cumbersome. To still be able to move and enjoy larger environment, alternative locomotion techniques have been developed. The work of Martinez et al. [61] have categorized these techniques into five major categories: Walking-based, steering-based, selection-based, manipulation-based and automated locomotion. Walking based techniques mostly rely on the physical movement of the body [9]. Here different techniques like walking in place [30], arm-swinging [14] or the use of external devices like omnidirectional treadmills [9] are very commonly used. The second most researched locomotion category [61] is steering-based locomotion. To navigate within a digital environment a continues movement either absolute or relative to the user is applied. This also creates the effect of floating or hovering within the digital space. These techniques are often combined with the input of a controller to determine the movement direction [87]. Alternative steering-based locomotion are head-directed [77], hand-directed [58] and lean-directed steering [11, 12]. For each technique the respective body part is directed in the desired movement direction. The research within the selection-based locomotion category is strongly dominated by different types of teleportation approaches. Here the user is provided with different types of inputs to select the desired location either by looking [37] or pointing [34] towards the destination. In most VR application the teleportation approach is very frequently employed, as this seems to be one of the most convenient way to move within the VR environment [9]. Manipulation based locomotion techniques are very unique approaches to solve the problem of limited physical space. These techniques manually manipulate the location of the user by virtual hand techniques. The user can grab the camera [79] or even the entire environment [21] with their hand. Subsequently moving their hand forward would pull the entire scene forward. The automated locomotion methods are usually based on predefined paths. The users themselves have little influence on the way this locomotion is carried out [81]. They can merely change the time at which locomotion is started or stopped and adjust the speed [72].

2.2 Hands-Free Interaction in VR

Hands-free interactions are necessary in special circumstances, such as medical scenarios under sterile conditions [23, 64]. Monteiro et al. [67] provide an overview in which they identify the main hands-free interaction techniques, the main interaction tasks that are tackled, and the metrics currently used for them. The authors identify *voice, eye* and *head* as the most commonly studied interaction modalities. Voice-based interaction studies can generally be divided into two main categories. Firstly, systems that use simple one-worded voice commands [17, 89] such as 'open' or 'close', and secondly, systems that can recognize and process complete sentences [1, 60]. Here, the user is usually given some kind of real-time feedback. Alternatively, speech can be used to fill in some textual properties [75] by dictating the desired text by voice. Voice-based systems do not necessarily have to be speech-based, but can also respond to certain sounds [94, 110]. For example Zielasko et al. [110] used the sound of a whistle as a start or stop command. This has reduced some possible complications and addressed the problem of multilingual speech processing. Eye tracking is also a popular method for hands-free interactions in Virtual environments [67]. While it can be difficult to add eye tracking to an already established HMD, some newer HMDs already support internal eye tracking. By tracking the eye position in real time, the user can select or point to a virtual item without moving their head [8]. It is also possible to track eye gestures such as blinking [49] or closing the eyes [44] to confirm a selection. However, using the eyes as a tool to interact with a virtual environment can also be problematic. Since people naturally move their eyes or blink, this can easily be misinterpreted as an unwanted interaction command [40]. In addition to eye tracking, head tracking can also be used to process input data. A simple nod or shake of the head can be used to confirm or deny a particular interaction [73]. Placement of an indicator that moves relative to the head position could be used to uniquely select specific items in the digital world by holding the indicator over the item for a period of time [103, 104]. Other less common alternatives for hands-free interaction include foot tracking [65], brain activity tracking [73], and body tracking [33].

2.3 Voice Interaction Systems in VR

Speech interfaces in VR applications come in different varieties. A predominant use of voice interaction is found in *command interfaces*. A fixed set of commands is accessible to the user and can be articulated via natural language. When the system perceives speech, it analyzes it and triggers the intended action, such as in voice-controlled positioning of virtual implants in a surgical planning process [82]. To benefit from immersive data visualization, voice interaction is a suitable tool, as speech is much more intuitive and quicker to apply here than keyboard input [39, 50]. In terms of locomotion, approaches combining the modalities of speech with gestures, head movements and gaze have been explored [35, 93],

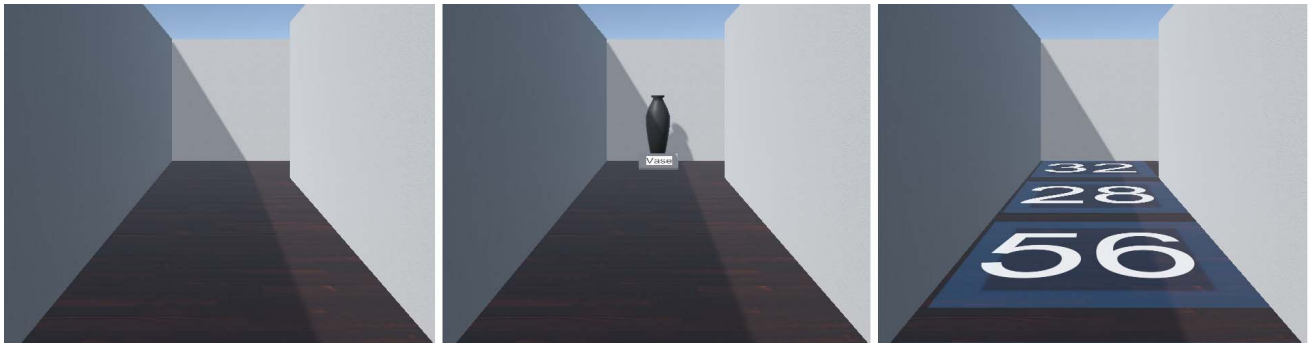

Figure 1: *Left:* Steering based locomotion techniques relying on voice commands or leaning. *Center:* Landmark-based teleportation using voice to articulate visual landmarks. *Right:* Number grid-based teleportation using numbers as an alternative in situations where no articulable landmarks are present.

finding that combining those modalities increases the presence of users in the VR experience. *Dialogue* interfaces find application in interaction with virtual avatars and agents [2, 96]. Morotti et al. [69] used a voice assistant in a virtual reality fashion shopping experience. Chilufya and Arvola [20] design a virtual receptionist that provides information via speech in different rooms of a university building. The influence and interaction of speech and visual elements on the narrative in a VR experience is explored by Osking et. al. [70], who found that the use of voice control in a VR experience increases the emotional impact on the user. Voice and sound provide an alternative channel that enables the inclusion of people and helps them participate in a virtual experience, such as in Ferracani et al. [31] making museums navigable for people with motor disabilities. Our work is related to this in that it provides alternative channels of interaction for scenarios where hands are not available.

2.4 Speech Technologies in VR

Speech interfaces in VR are based on work in the field of NLP. The entire speech pipeline first involves the conversion of audio signals into text strings, which are then analyzed for their semantic content. Text responses are then generated based on this and communicated to the user via speech synthesis. Advances in speech-to-text systems in recent years stem from successful deep learning approaches. *Wav2vec* [6, 88] as a key idea uses large-scale pre-training on raw audio data to learn discrete representations of audio segments, which are then matched to transcribed speech in a second step. There are several variations of this approach, using noisy student training [108] or quantization techniques [5] to improve performance. *Kaldi* [76, 84] and *VOSK* [92] provide performance-optimized approaches for use in real-time scenarios. In our implementation, we use a *VOSK* model for speech recognition because these models are small and work fast [47]. To reliably analyze the semantic content of an utterance, natural language understanding methods use large language models such as *BERT* [25] or *XLNET* [105]. On the one hand, these approaches offer powerful semantic disambiguation capabilities, but on the other hand, they suffer from long inference times due to their model size. Faster inference can be achieved by knowledge distillation as in *DistilBERT* [85]. For hard real-time requirements, grammar-based semantic parsing approaches or approaches based on rule- or regular expression matching lead to a faster inference [13, 54, 86]. In practice, practitioners have to trade off between speed and accuracy, depending on the requirements of the application at hand. Since real-time capability plays an important role in our experiments and we measure the time to target in our participants' walks, we opted for a matching-based semantic parsing approach in our implementation.

3 Methods

In this section, we explain the three different speech-based methods of hands-free locomotion and the leaning-based locomotion technique. We explain why they were chosen and discuss details of their implementation. Section 3.1 deals with the two steering-based techniques which are voice-based steering (3.1.1) and leaning (3.1.2). In Section 3.2 we delve into selection-based methods by focusing on landmark-based teleportation (3.2.1) as well as number grid teleportation (3.2.2) as an elegant solution for environments where landmarks are unavailable or difficult to articulate (as e.g. certain medical terms in virtual medical environments).

3.1 Steering-based Locomotion

Steering-based locomotion techniques are locomotion techniques that allow the user to control the direction and speed of movement in real time [57]. For steering-based locomotion, we choose to compare voice-based steering with leaning-based steering, as both are applicable in hands-free scenarios. Both techniques allow the user to change the direction and speed of a continuous movement via hands-free control signals.

3.1.1 Voice-based Steering

Voice-based steering allows users to change direction and speed through a series of voice commands. To change the direction, users can choose between the direction commands `<forward>`,`<backward>`,`<left>`,`<right>` and `<start>` and `<stop>` commands. Once the command is executed, the user's movement is adjusted based on their current view direction. For example, if the command `<left>` is applied, the user will be continuously moved towards the left of their current view. This applies to all other commands as well, with their corresponding direction. By modifying the current movement command, such as switching from `<left>` to `<right>`, or by altering the view direction, the global movement direction can be changed. It is worth noting that the voice commands only control the movement direction of the digital avatar and do not affect its rotation. In order for the user to turn within the virtual environment, they must also physically rotate their body in the corresponding direction. To change the speed of a movement, users can choose between the commands `<faster>` or `<slower>` to gradually increase or decrease the speed. Once a voice command has been successfully parsed, the direction or speed changes immediately. Between commands, the speed of the movement remains constant. For a detailed explanation of the voice control implementation, see Section 3.3.

3.1.2 Leaning

In leaning-based locomotion, users specify the direction and speed of movement by leaning their heads in different directions. To avoid unwanted movements, the user has an area in which he/she can move freely and in which no external movements take place. This area has a radius of 25 cm around the original center of the head. In order for the user to know at what distance the movement starts, a small circle is displayed on the floor. As soon as the user's head moves outside this area, locomotion is initiated. By creating this move-free area, the user can continue to physically move, rotate, and reposition without triggering unwanted movement in the digital environment. Once the user leans outside of this area, their position is smoothly moved in the leaning direction. The speed of locomotion is controlled by distance. The further the user leans in a certain direction, the faster the movement. As soon as the user returns to an upright position, the movement is completely stopped. The system responds to user head movements in real-time and immediately adjusts direction and speed. The locomotion works best when the user's feet are also tracked. This allows the center position and leaning range to be automatically updated in real-time. However, as long as the user's physical location does not change, as in our study, foot tracking is not required and therefore not utilized in our study. The initial position can be calibrated when the application is started. As soon as the user intends to change his/her physical position, the leaning approach can be stopped using the <stop> voice command. Once the new physical position is reached, the leaning technique can be restarted with the <start> command. Whenever the approach is restarted, the center point and the leaning area are recalibrated.

3.2 Selection-based Locomotion

Selection-based locomotion techniques allow the user to select a target location in an environment and then move to that position either immediately, for example via teleportation, or in a transitive movement [57]. For the evaluation of selection-based locomotion techniques, we have chosen to compare landmark-based teleportation and number grid-based teleportation for the following reasons: (1) Gaze and eye tracking methods have been presented as alternative approaches for selection-based locomotion in hands-free scenarios [68]. Existing work by Calandra et al. [15] provides a comprehensive comparison of landmark selection using speech and gaze in point-to-point navigation tasks, concluding that gaze combined with speech and point-of-interest descriptions leads to higher accuracy in identifying the navigation target. Based on these results, we did not compare landmark selection by speech and gaze a second time and refer to the results of Calandra et al. [15]. (2) Instead, we found that selection-based locomotion methods are highly dependent on the quality, identifiability, and articulability of landmarks in the VR scene. This issue is interesting because challenging virtual environments, such as immersive medical visualizations [29] or abstract data visualizations [95], may have no landmarks or landmarks that are difficult to articulate. This led us to compare speech-based landmark teleportation with number grids, which may be a potential solution for speech-based, hands-free locomotion.

3.2.1 Landmark Teleportation

Landmark teleportation allows the user to describe a visual landmark in natural language, which is then located by the system. To ensure that the user does not teleport directly into a 3D object, the target location is determined based on the center of the landmark and an offset distance. This offset distance is calculated using the bounding sphere of the mesh and ensures that the target location is a few centimeters away from the landmark, in the direction of the user. This helps prevent any potential issues with teleportation. To specify a teleportation, users can use the commands <teleport> or <jump>. Landmarks are referenced by their linguistic descriptions. In our study scenario

Figure 2: The ten visual landmarks <chair>, <mirror>, <lamp>, <cupboard>, <vase>, <table>, <bed>, <sink>, <kitchen> and <sofa> that are included in the user study.

(see Section 4), the following ten objects are used as visual landmarks: <chair>, <mirror>, <lamp>, <cupboard>, <vase>, <table>, <bed>, <sink>, <kitchen>, <sofa>. A typical command for a landmark teleportation looks like this: 'jump to the bed'. When a teleportation command is received, the system performs an immediate transformation of the location to the position in front of the destination. The technical implementation of speech recognition is identical to that described in Section 3.3. During our user study, (see Section 4) landmarks are positioned randomly.

3.2.2 Number Grid Teleportation

In number grid teleportation, an ad hoc grid is created as overlay on the floor and filled with random numbers. The grid metaphor is inspired by the coordinates on a *chessboard*, which makes parts of an object that has no visually distinguishable landmarks uniquely identifiable and navigable via natural language. For our application, the world is procedurally divided into square areas. Each of these squares is then represented by a unique number that can be used for teleportation. The size of the grid and thus the number of possible teleportation locations can be changed to achieve an adjustable density of teleportation points. Because the path used in our study is very narrow, the numbers are created in only a single column. However, for wider areas, they can be divided into as many rows and columns as necessary. By sorting these numbers in any pattern, e.g. ascending from start to finish, the user can quickly jump to the beginning, middle, or end of a path without knowing where these locations are. To guarantee independent results and prevent study participants from memorizing number positions on the grid, the numbers in the grid are randomly generated during the experiments (see Section 4). Target positions on the grid are referenced by users articulating the nearest number to the desired target position. For better auditory discrimination, we limit the set of numbers to numbers between <eleven> and <ninety-nine>. A typical command for a number grid teleportation looks like this: 'teleport to fifteen'. Similar to 3.2.1, upon receiving a command, the system transforms the user's location into the grid cell containing the articulated number immediately. While we have chosen numbers to uniquely identify each cell, theoretically any type of description can be used.

3.3 Voice Control

Our voice control system consists of a speech recognition part and a semantic parsing part. The speech recognition component is based on a speech-to-text service implemented as a Python server [97]. The server runs in the background as a Docker [63] container. Our Unity [36] frontend application communicates with this service via HTTP web requests. To realize speech recognition, we use the microphone of a *Valve Index* head-mounted display (HMD). The audio

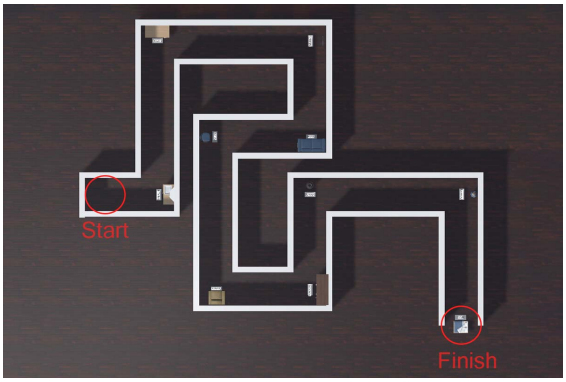

Figure 3: Example path with nine turns containing all ten locations needed for landmark-based teleportation.

stream is recorded at a sampling rate of 16kHz and simultaneously streamed to the speech recognition service. A permanent transcription function running in the backend translates the raw audio blocks into text strings using the offline speech recognition model *vosk-small-en-us* by *VOSK* [92]. The open source project *VOSK* provides recognition models with different complexities and accuracies. All implementations are based on the *KALDI* speech toolkit [76]. We chose the smallest speech model to reduce latency so that users feel comfortable with speech interaction. Since we deploy the voice control system as well as the VR system on the same machine and the GPU is required for rendering the VR environment, the speech recognition service must use CPU resources only. With this combination, we ensure the lowest possible latencies for our users. The *VOSK* speech recognition model outputs a substring for each recognized word and a final string for a fully recognized utterance. Both the partial strings and the final strings are immediately returned to the Unity frontend application. The string is then parsed in real time using regular expressions. For this purpose, a fixed vocabulary is defined consisting of the above-mentioned `commands` and `landmarks`. Partial strings are immediately displayed to the user, providing immediate feedback on what the system has understood. The final strings arriving at the Unity front end are passed through the regular expression matcher, which looks for commands and landmarks. Each parsing loop ends with a list of matches, which are used to populate the `command` and `landmark`/ or `number` slots in a provided locomotion function. If all slots are filled, the locomotion function is executed and the locomotion takes place immediately. If no match is found, the user is prompted to try again.

4 User Study

Satisfaction, efficiency, and efficacy are the most used evaluation metrics in studies to assess the usability of interfaces. Usually, they are acquired through custom questionnaires that gather feedback from users about their preferences and use of the interface. A few studies rely on validated questionnaires [12, 61, 67].

4.1 Study Setup

The user study was conducted as an in-person, in-lab study, with the same equipment provided to each participant. This way we can ensure that the conditions are comparable for all participants. Since we had to process each participant's voice, the lab was in an isolated room. No one except the participant was allowed to speak during the study. After welcoming the participants and introducing them to the research topic, we immediately placed them into the virtual world. Here, they were presented with a demo scene in which they could learn and freely explore a given locomotion technique. After all ambiguities and questions from the participants were resolved, the actual study was started. We additionally explained that questions during the study were only allowed in extreme cases where users were unable to fulfill the task, as otherwise the verbal input could be interpreted as an unwanted locomotion command. Once the participant finished the study for the first locomotion technique, he/she was placed in the next demo scene. From here, the next locomotion technique was explained and explored, and subsequently evaluated. This process was repeated until each participant had performed all four locomotion techniques. To eliminate possible order effects, locomotion techniques were presented to each participant in counterbalanced fashion. After the study, each participant engaged in a questionnaire that included demographic questions as well as questions about discomfort, comfort, presence, and usability. After completing the questionnaire, a short debriefing session was carried out in which additional feedback on the study could be provided. Completing the study, including all necessary preparations, the study itself and the post study questionnaire took approximately 30 minutes per participant. All measures used in the study and the questionnaire are explained in Section 4.4.

4.2 Hardware and Software Setup

The virtual test environment for the study needed to be developed in a development framework capable of building applications for VR. We chose Unity [36] as the main platform for designing our environment because it delivers high compatibility with a wide range of devices, can be developed and deployed on different operating systems, and is supported by a strong community in case of bugs and problems. For the final study, we chose a desktop PC with an AMD R9 5900x CPU @ 3.7 GHz, 32 GB of RAM, and the Nvidia GeForce RTX 3080ti. This setup allowed us to run both the VR application and the speech processing as stable as possible. For the head-mounted display, we chose the *Valve Index* due to its built-in high-quality microphone. Since our approach consists of evaluating hands-free locomotion techniques, we did not use controllers or hand tracking within this setup. The only way to communicate with the system was by voice. Considering that the immersion and realism of the application could be affected by the 3D models in the scene, we decided to use realistic furniture to create a coherent environment. However, since visual representation is secondary in our case, we had to make some practical adjustments given the object size. Some life-size landmarks, such as <`bed`> or <`kitchen`>, would take up too much space in our environment. We scaled them to fit better into the given environment. To still achieve a realistic look, we chose a high-quality wallpaper texture as well as a high-quality wood texture for the floor, shown in Figure 2.

4.3 Procedure

The study itself was divided into three different phases: (1) preparation phase, (2) first trial, (3) second trial. During the preparation phase, each participant had time to get used to the different locomotion techniques within an separate digital room. This room contains only the necessary landmarks or numbers and instructions on how to use the locomotion technique. To ensure that all participants feel comfortable and confident with each locomotion technique, we did not impose any time limits on their practice. This approach allows participants to fully understand and master the techniques, regardless of their familiarity with VR. During the training, we presented the details of the upcoming study so that each participant knew what to do once the study began. When the participants were completely sure that they had fully understood both the locomotion technique and the goal of the task to be solved, the experiment was started. During the first trial, the goal for each participant was to navigate through a simplified maze-like structure. This linear path had only branches to the left and right, but no dead ends. In this way, reaching the end of the path is not determined by chance. In total, the predefined path had nine turns with ten different landmarks (nine turns +

one at the finish), as shown in Figure 3. Once the participant reached the end of the , a new path was generated and the procedure was performed again. Both mazes were constructed so that the length and number of turns were exactly the same. Although the velocity of the movement can have an influence on the temporal performance, the teleportation and the timing by movement from one landmark to the next are approximately similar. By including a second trial with a different pathway, we made it possible to observe whether the locomotion technique produced an improvement once the user became more familiar with the locomotion technique. While we want to investigate the learning effect within each locomotion technique, we also want to counterbalance the learning effect caused by the experimental setup itself. Therefore, we made sure that the order in which the locomotion techniques are performed varies for each participant. Although we recognize that the results of the study may not be fully stabilized after just two trials, we also understand that adding more trials could significantly increase participant fatigue and potentially impact the results in various ways. Despite this, we believe that initial changes and trends should still be observable with a relatively small number of trials. Since some locomotion techniques require landmarks, we placed them at each possible turn in the path. For landmark teleportation, each turn was randomly assigned one of the items presented in Section 3.2.1. For number grid-based teleportation, the floor was divided into different areas, with a random number assigned to each area (see Figure 1). For the steering-based locomotion techniques, the path consisted only of walls and the floor.

4.4 Measures

In order to appropriately evaluate the study conducted, we need to consider several aspects of our locomotion methods. The most common way to evaluate these methods against each other is to measure the time to target [12,61,62]. However, the time to complete a particular task is not the only important metric. In addition to temporal performance, the feeling provided by a particular method of locomotion may be even more important to some users, as this directly contributes to the immersion and usability of the technique in question. Therefore, we evaluate time as an objective measurement and include four subjective measurements. The above measures were chosen because they have already provided useful insights in various studies on locomotion [12, 61]. In addition, each measure is based on previously evaluated questionnaires. In the next section, we will discuss the announced measures of *performance, sickness, comfort, presence*, and *usability*.

4.4.1 Performance

To evaluate the performance of our system, we incorporate two metrics that are tracked during the study. The time-to-target measure is used as a performance measure by counting the number of seconds it takes a user to complete the path using the given hands-free locomotion technique. This is a well-known technique used in most locomotion studies [12, 61, 62]. In addition to temporal performance, we also want to evaluate the accuracy of our speech processing. Therefore, we perform a measurement that counts both the accepted speech commands and the failed commands. In this way, we also gain insight into how many speech commands are required to complete the path.

4.4.2 Sickness

Developing the most accurate and fastest locomotion technique would be pointless if the user would not be able to use it due to motion sickness or other discomforts. Therefore, we evaluate different types of discomfort by including four questions from the Simulator Sickness Questionnaire (SSQ) [42]: (a) *general discomfort*, (b) *headache*, (c) *eye strain*, and (d) *nausea*. We carefully selected these four questions (1, 3, 4 and 8) of the SSQ , as they have been identified as essential in similar research [12]. This way we can ensure that we obtained the most valuable information while minimizing participant burden. In a post-study questionnaire participants rated how severe each discomfort was on a 4-point Likert scale: (1) *none*, (2) *mild*, (3) *moderate*, (4) *severe*. Rather than using a pre- and post-study questionnaire to assess differences in discomfort, we worded the questionnaire to directly address changes in discomfort rather than the overall level of discomfort. In this way, we can also exclude a possible negative influence of the SSQ questions before the test, which might affect the results of the questionnaire after the study [107].

4.4.3 Comfort

While discomfort is a crucial exclusion criterion for locomotion techniques, comfort also plays an important role in the immersion and usability of locomotion techniques. The greater the comfort of using a particular locomotion technique, the higher the likelihood that this technique will be used again. To assess the comfort of locomotion techniques, we adapted the Device Assessment Questionnaire (DAS) [27]. From the DAS, we included only questions relevant to our work (2, 3, 4, 5, 6) about overall comfort, ease of use, accuracy, and physical and mental exertion. We removed the question about fatigue of body parts because we mainly evaluate speech-based locomotion. The questions can be answered on a 5-point Likert scale: (1) *strongly disagree* - (5) *strongly agree*. While the general goal is to develop a locomotion technique that has both a high degree of comfort and a high degree of accuracy, in most cases we can trade some accuracy for comfort. Whether precision or comfort is preferred depends primarily on user preferences and application.

4.4.4 Presence

Presence or immersion describes the extent to which the user has the feeling of 'being' in the digital world. To measure participants' sense of presence in the virtual environment, we adapted the Igroup Presence Questionnaire (IPQ) [90] to the needs of our study by reducing the number of items from 14 to 10. This decision was made to minimize participant distress while ensuring that essential information weres still captured. The deleted questions were deemed redundant or not critical to our research objectives. The remaining 10 questions were grouped into the four categories defined in the original IPQ questionnaire: general presence (1), spatial presence (2, 5, 6), involvement (8, 9, 10), and experienced realism (11, 12, 13), with multiple questions for each category. Each question can be answered on a 5-point Likert scale representing the degree of immersion: (1) *not present at all* - (5) *completely present*. Presence has a variety of influencing factors, such as realistic graphics, the field of view, and many others that cannot be considered in this study. Therefore, we are more interested in whether there are differences between the different techniques rather than the overall rating. Since all techniques are performed in the same environment, changes in presence are directly related to the locomotion.

4.4.5 Usability

To evaluate usability, we chose the System Usability Scale (SUS) [10], which contains ten questions. Each question can be answered on a 5-point Likert scale: (1) *strongly disagree* - (5) *strongly agree*. To gain a comprehensive understanding of the usability of the system, we used the ten questions of the SUS and divided them into categories relevant to our study. These categories were chosen to organize and analyze the data in a way that was both meaningful and consistent with the goals of our study. The categories are: (a) *Reusability* (1) - How willing are participants to use the system in the future? (b) *Complexity* (2, 3, 4) - How difficult is the given system to understand? (c) *Integration* (5) - How well is the locomotion integrated into the system? (d) *Consistency* (6) - How reliable does

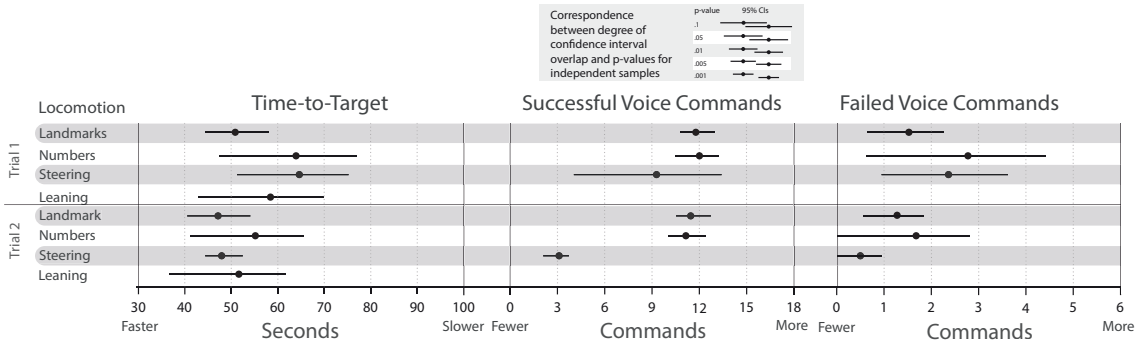

Figure 4: Measurement of time to target, successful / failed voice commands for landmark, number, steering, and leaning-based locomotion methods. The results are presented in a 95% confidence interval. The part on the correspondence of confidence intervals and p-values is based on [48].

the system work? (e) *Learnability* (7, 10) - How easy is it to perform a simple task when confronted with the environment for the first time? (f) *Confidence* (8, 9) - How confident is the user in operating the system. Each question is posed so that a high score for each category represents a positive result for overall usability.

4.5 Participants

For our study, we recruited a variety of participants from different backgrounds. There was no participant restriction for the study. Thus, we were able to recruit a total of 20 participants, of which 11 were male, 9 were female, and 0 were diverse. The ages of all participants ranged from 20 to 61, with an median age of 27. All participants had at least a high school diploma, with 11 of them currently enrolled in a bachelor's (3) or master's (8) degree program. Non-students could be categorized as salespeople (1), PhD candidates (2), physicians (4), professors (2), and others. While most participants (13) had at least some knowledge of 3D environments, 9 participants reported an interest in VR for more than a year. 13 participants have used VR at least once or twice in their lives, and 2 use it almost every week. Of the 17 participants who use a PC daily, 11 play video games at least a few times a month and are familiar with different ways of locomotion in digital environments.

5 Results

To avoid dichotomous thinking [28], we will report our results as confidence intervals rather than p-values as recommended by the APA [3]. This approach is increasingly used across a variety of studies within the HCI [22] and VR community [41, 66, 102]. Our performance results (see Figure 4) represent an estimate of the bootstrapped confidence interval that includes the true mean 95% of the time. These results also include the effect size [24]. The questionnaire results, as shown in Figure 5 to 8, are presented with mean and standard deviation. For simplicity, we refer only to the mean values in the text. More detailed information can be found in the respective figures.

5.1 Performance

The performance measures, including *time to target* and *successful/failed voice commands*, can be found in Figure 4. The landmark approach (51.0s) showed the fastest average time-to-target, followed by leaning (58.4s), teleport to numbers (63.8s), and steering (64.6s) within the first trial. Although the average user was fastest with landmarks, those who were very skilled with leaning were the fastest users overall. In the second trial, we observed an improvement in time-to-target for all locomotion methods. Note that both paths had exactly the same length and number of turns. While landmarks were still the fastest (47.1 s), voice-steering followed very closely behind (47.6 s). Nevertheless, the overall fastest user still used the leaning approach. In addition to evaluating the temporal performance of each locomotion technique, we also performed a measurement of the voice processing based on the successful and failed voice commands. Since the teleportation techniques require at least 10 successful voice commands (one for each landmark) to reach the end of the path, it is not surprising that the highest successful voice commands are for teleportation by numbers (12.1) and teleportation by landmarks (11.9), followed by voice-steering (9.3). Since the leaning approach does not require a voice command, we do not have a measurement for successful or failed voice commands. While the number of successful voice commands required to reach the end of the second path decreases slightly for teleportation through landmarks (11.4) and numbers (11.1), the number of successful commands required for voice-steering decreased dramatically to 3.1. When considering the failed speech commands, it should be noted that any type of speech that could not be mapped to one of the locomotion commands is considered a 'failed' command. Therefore, this measure should be taken with caution, as some participants still asked questions or used speech to express pleasure or frustration, which also counted as a failed voice command. However, we attempted to minimize this false error by frequently reminding participants not to speak during the study. However, failed voice commands based on incorrect processing of speech are also included. With an average of 2.1 failed speech commands for the first path, we can see a clear improvement for the second path with 1.1 failed commands.

5.2 Sickness

The results for sickness can be found in Figure 5. While the results show that the scores for overall sickness, headache, eye strain, and nausea are very low overall, leaning caused a small increase in overall sickness and nausea compared to the other locomotion techniques. Looking more closely at the underlying data, we found that most participants had no problems with leaning. However, those who did experience sickness reported a higher score of about 3 (moderate sickness), making leaning largely unusable for them. Although we cannot determine the long-term impact of various techniques on motion sickness, we have noticed that if motion sickness is present, it tends to appear early in the first trial. A minimal increase in eye strain was observed for both teleportation techniques compared to the steering techniques. Although there are small differences between sickness scores, the average person does not experience large sickness differences between the investigated locomotion methods.

5.3 Comfort

The results for comfort can be found in Figure 6, which includes ratings for ease of use, mental effort, physical effort, accuracy, and

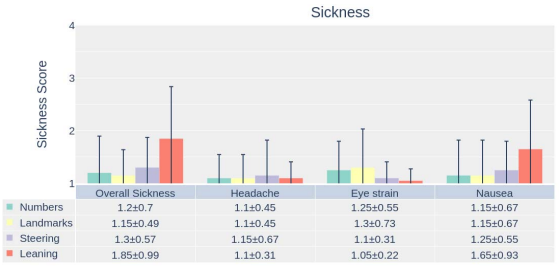

Figure 5: Mean scores of sickness measures. Capped vertical bars indicate ± SE. Leaning shows higher sickness measures than voice-based steering and teleportation techniques.

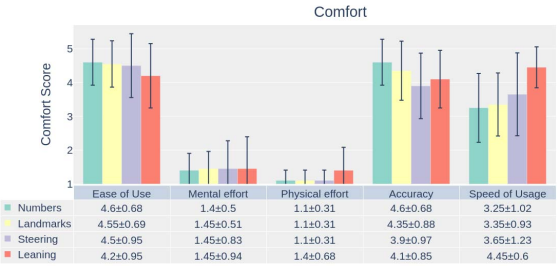

Figure 6: Mean scores of comfort measures. Capped vertical bars indicate ± SE. Landmark and number teleportation have higher accuracy and ease of use, while leaning has the highest speed of use.

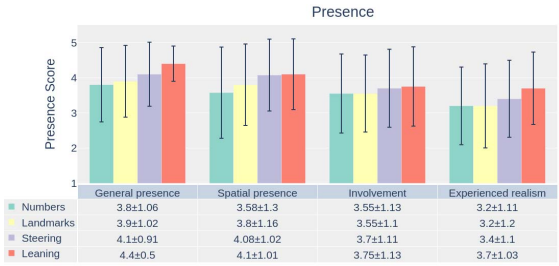

Figure 7: Means of general presence, spatial presence, involvement and experienced realism. Capped vertical bars indicate ± SE. Steering methods show higher scores than selection methods.

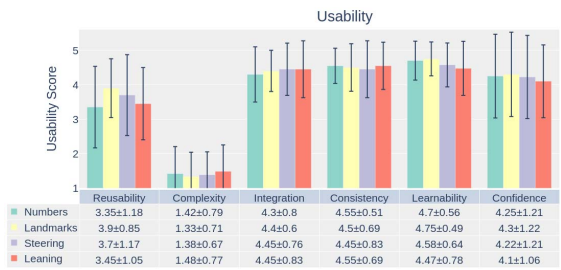

Figure 8: Mean scores of usability measures. Capped vertical bars indicate ± SE. Voice-based techniques show similar results to leaning as an established alternative.

speed of usage. While all locomotion methods generally have a high rating for ease of use (4.2-4.6), the leaning method has a minimally lower rating (4.2). Mental (1.4 - 1.45) and physical (1.1 - 1.4) effort is very low for all locomotion methods, with leaning having a slightly higher physical effort score (1.4). This was to be expected as leaning was the only technique that required the participant to physically move their body to navigate the environment. The accuracy of all techniques ranged from 3.9 to 4.6, with teleporting to numbers (4.6) and teleporting to landmarks (4.35) having slightly higher accuracy than leaning (4.1) and steering (3.9). The speed of usage score shows that the locomotion techniques with a high accuracy take more time to execute. The speed of usage score is highest for leaning (4.45), followed by steering (3.65), teleporting to landmarks (3.35), and teleporting to numbers (3.25).

5.4 Presence

All results for the presence measure can be found in Figure 7. Overall presence is highest for leaning (4.4), followed by steering (4.1), teleportation through landmarks (3.9), and teleportation through numbers (3.8). The same order is found for spatial presence: leaning (4.1), steering (4.08), teleportation by landmarks (3.8), and teleportation by numbers (3.58). For involvement in the virtual environment, the scores of all locomotion techniques are very close (3.55-3.75). The results show that for the leaning technique (3.7) the experienced realism is higher than for steering (3.4), teleportation by landmarks (3.2) and teleportation by numbers (3.2).

5.5 Usability

The results of the usability questionnaire (see Figure 8) show that most of the values for reusability, complexity, integration, consistency, learnability and confidence are similar for all locomotion methods. The largest difference is evident in the reusability score (3.35-3.9). Here, the landmark approach performed best (3.9), followed by steering (3.7), leaning (3.45), and teleportation by numbers (3.35). The complexity of each location method was very low, with an average score between 1.33 and 1.48. The integration into the system was also rated very similarly for all techniques (4.3-4.55). The same was true for consistency (4.45-4.55), learnability (4.47-4.75), and confidence (4.1-4.3).

6 Discussion

Our results produced interesting comparisons in terms of control using speech as well as revealed failure cases that we will look at in more detail in the following. One of the more notable characteristics of the *performance* results is that the time to complete the path decreased dramatically for the second path. This clearly shows that the user needs a little time to get used to each locomotion technique in order to unlock its real potential. While the teleportation approach was fastest on average, we also found that the overall fastest results were obtained for the leaning approach. This performance may be directly related to the overall sickness and nausea scores. Participants who did not experience motion sickness during the leaning method generally performed very well with this technique. However, participants who experienced even a little motion sickness reported that leaning was the most uncomfortable way to navigate. This was also evident in the time-to-target measurement. These participants took longer to navigate the path with leaning than with any other method of locomotion. Because the performance of this technique is highly dependent on each user's preferences and tolerance for motion sickness, we would only recommend using this technique if the user is confident in their tolerance for motion sickness. We also believe that the slightly lower score in the other categories is mainly due to motion sickness, as this was one of the most common responses during our debriefing. While motion sickness was not unique to the leaning approach, some participants experienced problems when using the voice-steering approach. Since both approaches involve floating motion, motion sickness is likely not directly related to the input method of leaning, but to the way motion is applied.

We also found that voice-steering received the most positive feedback while also being the method that could be improved the

most. Once participants realized that they did not have to stop, turn around, and restart the movement at every turn, the number of voice commands processed was drastically reduced. Instead of stopping at every turn, participants realized that they could simply 'look around the corner' since the forward direction of the movement was tied to their viewing direction. This can be clearly seen in the successful voice commands in Figure 5. About halfway through the first experiment, most participants realized how practical the method actually was. This conclusion is confirmed by the drastically reduced number of voice commands for the second path. Participants generally used only the forward command at the beginning of the path and then did not need to give another command until they reach the end. This also explains why steering by voice resulted in the strongest improvement in time-to-target measures between the two paths. During our feedback session, participants also mentioned that the slight delay in processing the voice was cumbersome and needed to be improved in order to use the <stop> command properly. Otherwise, they would already be very close to a wall before the movement would actually stop. However, they felt that this method of navigating by voice could be very enjoyable and rated this method of locomotion as their favorite of all the techniques.

While voice-steering was voted the most popular method of locomotion, teleportation to the landmark came in second with only one less vote. This, along with the fact that the voice-steering approach needs some improvement, has probably led to the high reusability score of teleportation by landmark. Users were very confident in using this technique. This confidence was also reflected in the measurements of time-to-target. Here, teleportation to landmarks had the fastest results on average. One of the major drawbacks we found during the study was that it was difficult for non-native speakers to navigate using this technique. As the speech interface was only trained on very specific commands, participants' words were sometimes misunderstood, leading to frustration in some cases. However, since our speech processing can theoretically be trained on any language, we consider these cases as exceptions. In the future, these cases can be addressed by adapting the voice recognition model to specific dialects and intonations, as has been shown by [19, 83, 106].

Since numbers teleportation took a very similar approach to landmark teleportation, we were surprised that the time-to-target measurement was worse for the numbers approach. However, this could be explained by language processing, as numbers have additional problems with ambiguous number names. For example, the number <twenty-two> could also be interpreted as <twenty> and <two> if paused too long. We intentionally removed all numbers between zero and nine, as well as any increment of ten, to avoid false teleportation. However, the misinterpretation of the numbers as two separate numbers still resulted in a failed voice command. This problem seems to have occurred a few times during the study, as the number of failed voice commands is higher than for the landmark method. This also explains the slower time-to-target results. Since both approaches are based on voice command teleportation, we expect that both will perform equally well once the voice processing is optimized. However, both approaches have different use cases. The landmark approach can be used very easily in environments with many visual landmarks, while the number approach can be generated on any type of surface. This allows the user to teleport more accurately in less conceptually enumerable environments. A combination of the two could also be very beneficial. The user can use the landmark approach if landmarks are present, or create an ad hoc grid representation of the environment if needed. This way, the environment is not overloaded with numbers when they are not needed, which provides exciting ideas for future work.

6.1 Future Work and Limitations

Taking into account our user study, results, and feedback from our participants, we have identified some very interesting research topics to explore in the future. To get rid of a small, fixed vocabulary and enable a wide range of semantically variable commands, we plan to integrate more sophisticated language understanding models into our pipeline. This brings challenges related to the high inference times of these models, as we still want to provide speech services with consistently low latency. In this way, we will be able to support semantically similar commands that can be used in different locomotion approaches, making communication with the system more natural. By implementing a more advanced speech model, we have the ability to process arbitrary rotation and movement directions, allowing for a more intuitive way to steer.

In addition, we plan to include multilingual support to reach a wider audience and solve the problem of non-native speakers. Going forward, we plan to expand the functionality of our locomotion approach. We found that in most cases, the torso is still rotated when navigating around a corner. However, when the user interacts with real-world devices, such as a stationary haptic feedback device, locomotion around multiple corners is still difficult. Therefore, we want to incorporate additional voice commands that enable users to <turn> their virtual avatar or <circle> around a certain view. Our primary focus in this study was to explore the use of voice control for fundamental navigation. We therefore kept the design straightforward and did not include any complex scenarios. It remains to be seen how these voice-based techniques will perform in a more realistic, visually rich environment with multiple paths and potential occlusions. We hope to address this question in future work.

To further extend voice teleportation, we intend to combine the approach of number and landmark teleportation. As long as landmarks are present, we do not need to clutter the view with additional numbers. However, once the user approaches an area without sufficient landmarks, we plan to create an ad hoc grid system that extends relative to the user. In this way, the environment can be overlaid with additional landmarks to allow accurate teleportation via voice commands in almost any environment. How well this can be integrated into everyday VR and how best to create the grid representation is something we are excited to explore in our future research.

7 Conclusion

In this work, we investigated three speech-based, hands-free VR locomotion techniques and compared them to leaning as a proven alternative. Previous work indicated the usefulness of speech as an alternative locomotion method in hands-free scenarios by examining its effectiveness in combination with gaze movements. Our study extends this effort by focusing the comparison on leaning as another important alternative and examining the effects of speech-based locomotion techniques in environments where landmarks are not available or where they are difficult to articulate. Using the above-mentioned techniques, we conducted a user study with 20 participants to compare the objective performance of each locomotion technique and measure the subjective measures of presence, comfort, discomfort, and usability of them during a maze-like navigation experiment. Our analysis of the different variables revealed that users using voice-based teleportation to landmarks performed maze-like navigation tasks faster on average. Furthermore, speech-based techniques are indistinguishable from leaning in terms of presence and usability. In terms of nausea, leaning shows a higher risk of nausea, especially for participants with less VR experience. Interestingly, the results of landmark and number teleportation show similar values, for sickness, presence, comfort, and usability suggesting that in environments without or with hard-to-articulate landmarks, grid-based navigation with numbers is a good alternative. Based on these results, we conclude that voice-guided locomotion techniques are suitable for use in hands-free scenarios. This leaves room for a deeper investigation of voice-based ad hoc navigation techniques, as they are shown to be applicable regardless of the visual composition of the particular hands-free environment.

REFERENCES

[1] 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*, 9(7), 2018.

[2] G. Ali, H.-Q. Le, J. Kim, S. won Hwang, and J.-I. Hwang. Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable mixed reality environment. *32nd International Conference on Computer Animation and Social Agents*, 2019.

[3] American Psychological Association. *The Publication manual of the American psychological association*. Washington, DC, 6th ed., 2013.

[4] J. Austerjost, M. Porr, N. Riedel, D. U. 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*, 23:476 – 482, 2018.

[5] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *ArXiv*, abs/1910.05453, 2020.

[6] 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*, 33:12449–12460, 2020.

[7] 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 *CUI 2021-3rd Conference on Conversational User Interfaces*, pp. 1–9, 2021.

[8] 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*, pp. 1–9, 2018.

[9] D. A. Bowman, E. Kruijff, J. J. LaViola Jr, and I. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley, 2005.

[10] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[11] G. Bruder, V. Interrante, L. Phillips, and F. Steinicke. Redirecting walking and driving for natural navigation in immersive virtual environments. *IEEE transactions on visualization and cg*, 18(4):538–545, 2012.

[12] F. Buttussi and L. Chittaro. Locomotion in place in virtual reality: A comparative evaluation of joystick, teleport, and leaning. *IEEE transactions on visualization and computer graphics*, 27(1):125–136, 2019.

[13] Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 423–433, 2013.

[14] D. Calandra, F. Lamberti, and M. Migliorini. On the usability of consumer locomotion techniques in serious games: Comparing arm swinging, treadmills and walk-in-place. In *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*, pp. 348–352. IEEE, 2019.

[15] 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)*, pp. 85–90. IEEE, 2022.

[16] D. W. Carruth, C. R. Hudson, C. L. Bethel, M. Pleva, S. Ondás, and J. Juhár. Using hmd for immersive training of voice-based operation of small unmanned ground vehicles. In *HCI*, 2019.

[17] 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*, pp. 34–46. Springer, 2019.

[18] D. F. Castronovo. Design and development of a virtual reality educational game for archi- tectural and construction reviews. 2019.

[19] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and L. Seps. Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives. *Speech Communication*, 55(10):1033–1046, 2013.

[20] E. M. Chilufya and M. Arvola. Conceptual designing of a virtual receptionist: Remote desktop walkthrough and bodystorming in vr. *9th International Conference on Human-Agent Interaction*, 2021.

[21] I. Cho, J. Li, and Z. Wartell. Multi-scale 7dof view adjustment. *IEEE Transactions on Visualization and cg*, 24(3):1331–1344, 2017.

[22] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020.

[23] S. Cronin and G. Doherty. Touchless computer interfaces in hospitals: A review. *Health informatics journal*, 25(4):1325–1342, 2019.

[24] G. Cumming and S. Finch. Inference by eye: confidence intervals and how to read pictures of data. *American psychologist*, 60(2):170, 2005.

[25] 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.

[26] 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*, 32, 2019.

[27] S. A. Douglas, A. E. Kirkpatrick, and I. S. MacKenzie. Testing pointing device performance and user assessment with the iso 9241, part 9 standard. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 215–222, 1999.

[28] P. Dragicevic. Fair statistical communication in hci. In *Modern statistical methods for HCI*, pp. 291–330. Springer, 2016.

[29] P. Eulzer, M. Meuschke, G. Mistelbauer, and K. Lawonn. Vessel maps: A survey of map-like visualizations of the cardiovascular system. In *Computer Graphics Forum*, vol. 41, pp. 645–673, 2022.

[30] J. Feasel, M. C. Whitton, and J. D. Wendt. Llcm-wip: Low-latency, continuous-motion walking-in-place. In *2008 IEEE symposium on 3D user interfaces*, pp. 97–104. IEEE, 2008.

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

[32] M. Friedrich, S. Langer, and F. Frey. Combining gesture and voice control for mid-air manipulation of cad models in vr environments. *ArXiv*, abs/2011.09138, 2021.

[33] M. Gelsomini, G. Leonardi, and F. Garzotto. Embodied learning in immersive smart spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[34] S. I. Gray, J. Robertson, A. Manches, and G. Rajendran. Brainquest: The use of motivational design theories to create a cognitive training game supporting hot executive function. *International Journal of Human-Computer Studies*, 127:124–149, 2019.

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

[36] J. K. Haas. A history of the unity game engine. 2014.

[37] M. J. Habgood, D. Moore, D. Wilson, and S. Alapont. Rapid, continuous movement between nodes as an accessible virtual reality locomotion technique. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 371–378. IEEE, 2018.

[38] A. M. Hashemian and B. E. Riecke. Leaning-based 360° interfaces: Investigating virtual reality navigation interfaces with leaning-based-translation and full-rotation. In *HCI*, 2017.

[39] C. J. Hughes and J. Paton. Voice interaction for accessible immersive video players. In *VISIGRAPP*, 2021.

[40] 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)*, 9(2):152–169, 1991.

[41] M. J. Jung, J. S. Libaw, K. Ma, E. L. Whitlock, J. R. Feiner, and J. L. Sinskey. Pediatric distraction on induction of anesthesia with virtual reality and perioperative anxiolysis: a randomized controlled trial. *Anesthesia & Analgesia*, 132(3):798–806, 2021.

[42] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[43] R. Khan, J. E. Plahouras, B. C. Johnston, M. A. Scaffidi, S. C. Grover,

and C. M. Walsh. Virtual reality simulation training in endoscopy: a cochrane review and meta-analysis. *Endoscopy*, 2019.

[44] 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 *23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2017.

[45] 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*, 33:8067–8077, 2020.

[46] A. Kitson, A. M. Hashemian, E. R. Stepanova, E. Kruijff, and B. E. Riecke. Lean into it: Exploring leaning-based motion cueing interfaces for virtual reality movement. *2017 IEEE VR*, pp. 215–216, 2017.

[47] R. Kolobov, O. Okhapkina, O. Omelchishina, A. Platunov, R. Bedyakin, V. Moshkin, D. Menshikov, and N. Mikhaylovskiy. Mediaspeech: Multilanguage asr benchmark and dataset. *arXiv preprint arXiv:2103.16193*, 2021.

[48] M. Krzywinski and N. Altman. Error bars: the meaning of error bars is often misinterpreted, as is the statistical significance of their overlap. *Nature methods*, 10(10):921–923, 2013.

[49] 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)*, pp. 2358–2362. IEEE, 2016.

[50] M. Lange, J. Hjalmarsson, M. Cooper, A. Ynnerman, and V. Duong. 3d visualization and 3d and voice interaction in air traffic management. In *The Annual SIGRAD Conference. Special Theme-Real-Time Simulations. SIGRAD2003*, number 010, pp. 17–22. Citeseer, 2003.

[51] B. Lee, A. Srinivasan, J. T. Stasko, M. K. Tory, and V. Setlur. Multimodal interaction for data visualization. *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, 2018.

[52] 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.

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

[54] J. Li, M. Zhu, W. Lu, and G. Zhou. Improving semantic parsing with enriched synchronous context-free grammar. *TALLIP*, 16:1 – 24, 2015.

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

[56] R. Lohre, J. C. Wang, K.-U. Lewandrowski, and D. P. Goel. Virtual reality in spinal endoscopy: a paradigm shift in education to support spine surgeons. *Journal of spine surgery*, 6 Suppl 1:S208–S223, 2020.

[57] M. D. Luca, H. Seifi, S. Egan, and M. González-Franco. Locomotion vault: the extra mile in analyzing vr locomotion techniques. *CHI Conference on Human Factors in Computing Systems*, 2021.

[58] J.-L. Lugrin, A. Juchno, P. Schaper, M. Landeck, and M. E. Latoschik. Drone-steering: A novel vr traveling technique. In *25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2019.

[59] J.-L. Lugrin, F. Kern, R. Schmidt, C. Kleinbeck, D. Roth, C. Daxer, T. Feigl, C. Mutschler, and M. E. Latoschik. A location-based vr museum. *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pp. 1–2, 2018.

[60] 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*, vol. 11584. Springer, 2019.

[61] 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)*, pp. 270–280. IEEE, 2022.

[62] D. Medeiros, E. Cordeiro, D. Mendes, M. Sousa, A. Raposo, A. Ferreira, and J. Jorge. Effects of speed and transitions on target-based travel techniques. In *Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology*, pp. 327–328, 2016.

[63] D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[64] 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*, 12(2):291–305, 2017.

[65] 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*, pp. 1–9, 2019.

[66] D. Monteiro, H. Chen, H.-N. Liang, H. Tu, and H. Dub. Evaluating performance and gameplay of virtual reality sickness techniques in a first-person shooter game. In *2021 IEEE CoG*, pp. 1–8. IEEE, 2021.

[67] P. Monteiro, G. Goncalves, H. Coelho, M. Melo, and M. Bessa. Hands-free interaction in immersive virtual reality: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27:2702–2713, 5 2021. doi: 10.1109/TVCG.2021.3067687

[68] P. Monteiro, G. Gonçalves, H. Coelho, M. Melo, and M. Bessa. Hands-free interaction in immersive vr: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27:2702–2713, 2021.

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

[70] H. Osking and J. A. Doucette. Enhancing emotional effectiveness of virtual-reality experiences with voice control interfaces. In *iLRN*, 2019.

[71] D. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32:604–624, 2021.

[72] S. F. Paulo, D. Medeiros, P. B. Borges, J. Jorge, and D. S. Lopes. Improving camera travel for immersive colonography. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 748–749. IEEE, 2020.

[73] C. Pavlatos and V. Vita. Linguistic representation of power system signals. In *Electricity Distribution*, pp. 285–295. Springer, 2016.

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

[75] S. Pick, A. S. Puika, and T. W. Kuhlen. Swifter: Design and evaluation of a speech-based text input metaphor for immersive virtual environments. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 109–112. IEEE, 2016.

[76] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[77] Y. Y. Qian and R. J. Teather. Look to go: An empirical evaluation of eye-based travel in virtual reality. In *Proceedings of the Symposium on Spatial User Interaction*, pp. 130–140, 2018.

[78] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[79] M. Raees, S. Ullah, and S. U. Rahman. Ven-3dve: vision based egocentric navigation for 3d virtual environments. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 13(1):35–45, 2019.

[80] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[81] K. Rahimi, C. Banigan, and E. D. Ragan. Scene transitions and teleportation in virtual reality and the implications for spatial awareness and sickness. *IEEE transactions on visualization and computer graphics*, 26(6):2273–2287, 2018.

[82] H.-R. Rantamaa, J. Kangas, M. Jordan, H. Mehtonen, J. Mäkelä, K. Ronkainen, M. Turunen, O. Sundqvist, I. Syrjä, J. Järnstedt, and R. Raisamo. Evaluation of voice commands for mode change in virtual reality implant planning procedure. *International Journal of*

Computer Assisted Radiology and Surgery, 17:1981 – 1989, 2022.

[83] K. Rao and H. Sak. Multi-accent speech recognition with hierarchical grapheme based models. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4815–4819, 2017.

[84] 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)*, pp. 6465–6469. IEEE, 2019.

[85] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[86] A. Saparov, V. A. Saraswat, and T. Mitchell. A probabilistic generative grammar for semantic parsing. In *CoNLL*, 2017.

[87] S. P. Sargunam and E. D. Ragan. Evaluating joystick control for view rotation in virtual reality with continuous turning, discrete turning, and field-of-view reduction. In *Proceedings of the 3rd International Workshop on Interactive and Spatial Computing*, pp. 74–79, 2018.

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

[89] 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 *International conference on human-computer interaction*, pp. 54–61. Springer, 2017.

[90] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3):266–281, 2001.

[91] 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)*, pp. 4779–4783. IEEE, 2018.

[92] N. V. Shmyrev and other contributors. Vosk Speech Recognition Toolkit: Offline speech recognition API for An- droid, iOS, Raspberry Pi and servers with Python, Java, C and Node. `https://github.com/alphacep/vosk-api`, 2022.

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

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

[95] A. Sterzik, N. Lichtenberg, M. Krone, D. Cunningham, and K. Lawonn. Perceptual evaluation of common line variables for displaying uncertainty on molecular surfaces. 2022.

[96] 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. *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, 2:959–964, 2007.

[97] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[98] S. M. Ventura, F. Castronovo, D. Nikolic, and A. L. C. Ciribini. Implementation of virtual reality in construction education: a content-analysis based literature review. *J. Inf. Technol. Constr.*, 27:705–731, 2022.

[99] 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*, pp. 1–8, 2021.

[100] P. Wang, P. Wu, J. Wang, H.-L. Chi, and X. Wang. A critical review of the use of virtual reality in construction engineering education and training. *International Journal of Environmental Research and Public Health*, 15, 2018.

[101] 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.

[102] E. Wolf, N. Merdan, N. Dölinger, D. Mal, C. Wienrich, M. Botsch, and M. E. Latoschik. The embodiment of photorealistic avatars influences female body weight perception in virtual reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 65–74. IEEE, 2021.

[103] Y. Yan, Y. Shi, C. Yu, and Y. Shi. Headcross: Exploring head-based crossing selection on head-mounted displays. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

[104] Y. Yan, C. Yu, X. Yi, and Y. Shi. Headgesture: Hands-free input approach leveraging head movements for hmd devices. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–23, 2018.

[105] 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*, 32, 2019.

[106] S. Yoo, I. Song, and Y. Bengio. A highly adaptive acoustic model for accurate multi-dialect speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5716–5720. IEEE, 2019.

[107] S. D. Young, B. D. Adelstein, and S. R. Ellis. Demand characteristics of a questionnaire used to assess motion sickness in a virtual environment. In *IEEE virtual reality conference (VR 2006)*, pp. 97–102. IEEE, 2006.

[108] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *ArXiv*, abs/2010.10504, 2020.

[109] E. Zidianakis, N. Partarakis, S. Ntoa, A. Dimopoulos, S. Kopidaki, A. Ntagianta, E. Ntafotis, A. Xhako, Z. Pervolarakis, E. Kontaki, I. Zidianaki, A. Michelakis, M. Foukarakis, and C. Stephanidis. The invisible museum: A user-centric platform for creating virtual 3d exhibitions with vr support. *Electronics*, 2021.

[110] 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)*, pp. 40–49. IEEE, 2017.

[111] K. İncetan, I. O. Celik, A. N. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almalioglu, R. J. Chen, F. Mahmood, H. B. Gilbert, N. Durr, and M. Turan. Vr-caps: A virtual environment for capsule endoscopy. *Medical image analysis*, 70:101990, 2021.