Comparison of hands-free speech-based navigation techniques for virtual reality training

Comparison of Hands-Free Speech-Based Navigation Techniques for Virtual Reality Training

Davide Calandra
Dip. di Automatica e Informatica
Politecnico di Torino
Turin, Italy
davide.calandra@polito.it
0000-0003-0449-5752

Filippo Gabriele Pratticò
Dip. di Automatica e Informatica
Politecnico di Torino
Turin, Italy
filippogabriele.prattico@polito.it
0000-0001-7606-8552

Fabrizio Lamberti
Dip. di Automatica e Informatica
Politecnico di Torino
Turin, Italy
fabrizio.lamberti@polito.it
0000-0001-7703-1372

***Abstract*—When it comes to Virtual Reality (VR) training, the depicted scenarios can be characterized by a high level of complexity and extent. Speech-based interaction techniques can provide an intuitive, natural and effective way to navigate large Virtual Environments (VEs) without the need for handheld controllers, which may impair the execution of manual tasks or prevent the use of wearable haptic devices. In this study, three hands-free speech-based navigation techniques for VR, a speech-only technique, a speech with gaze variant (gaze to point to the destination, speech as trigger), and a combination of the fist two are compared by deploying them to a large VE representing a common industrial setting (a hangar). A within-subjects user study was carried out in order to assess the usability and the performance of the considered techniques.**

***Index Terms*—virtual reality, navigation, hands-free locomotion, speech recognition, training**

I. Introduction

Virtual Reality (VR) is becoming more and more widespread in many fields. Training, in particular, is one of the fields which are benefiting most from this technology. In this context, the use of VR allows the creation of arbitrarily wide Virtual Environments (VEs) capable to represent any training scenario. When the size of the Virtual Reality Training Scenario (VRTS) exceeds the free space available around the user in the real world, additional stationary navigation techniques [1] have to be supported to complement the room-scale natural locomotion.

To cope with this limitation, numerous works explored the possibility to exploit *Speech-based* interfaces to support a teleport-based navigation in VR scenarios [2]–[4]. The combined use of speech and gesture as input is not new [5], [6]. In the context of navigation in immersive VR, voice commands can be used alone to specify the desired POI to move to (e.g., by uttering its name) [2], or by combining phrases like "take me there" with a direction, which can be expressed via pointing gestures [3]. In order to support a completely *hands-free* approach, the need for a pointing gesture/handheld device can be avoided by exploiting the head-gaze orientation [4]. In some cases, voice commands may be considered as ambiguous by the system. In this case, additional disambiguation techniques may be required to manage ambiguous interactions [3], for example by providing additional information to the user to make the intent clear.

The aim of this work is to compare the performance of three different implementations for hands-free speech-based teleporting in VR when applied to a large indoor scenario representing a commonly industrial VTRS (i.e., a hangar [7]). In particular, a first speech-only technique, in which the user can pronounce phrases composed by the movement action and the name of the destination which have to be reached (e.g., "Take me to the yellow bin") [3], a second multi-modal technique combining the use of voice commands to trigger the teleport action and the head-gaze direction to indicate the desired destination [4], and a hybrid technique obtained by combining the functionalities of the two previous implementations. For each of them, an ad hoc disambiguation technique was also included.

II. Background

The navigation in large VEs for training purposes is a major open issue, as shown by the numerous of recent investigations on the topic [8], [9].

So far, large numbers of locomotion paradigms for VR have been proposed and widely investigated [10]–[12]. Among them, *Teleporting* represents one of the most popular approaches, being characterized by high intuitiveness, low cognitive demand and limited impact in terms of cybersickness [13]. For these reasons, it also widely adopted in VR training applications [14]. As many other techniques, teleporting relies on the use of the hand controllers commonly bundled with commercial VR systems. However this solution may not be ideal in case of training scenarios heavily based on hand interactions (e.g., manual tasks or manipulation of virtual objects), and it assumes that the user is able to hold a hand controller device, and this may not be true in case of training experiences involving the use of hand-tracking techniques [3] or wearable (buttonless) haptic devices [15].

As showed in [16], a wide number of ways to avoid the use of hands in VR has been proposed and studied in the years. Among them, the use of voice, from simple commands to Natural Language Processing (NLP), and eye/head gaze are the most investigated, followed by less common approaches such

This work has been partially funded by the VR@POLITO initiative.

as the use brain activity, facial expressions, foot movement, body position and contraction of muscles.

Being navigation a form of interaction, speech-based interaction can easily be applied for this purpose. For example, authors of [2] proposed a the use of voice commands to interact within an immersive experience in large VE (a museum). Among other functionalities, the speech can be used to move from one room to the other, effectively implementing a voice-based navigation system, with the aim to keep the application accessible to users with motor disabilities. Although the proposed paradigm guaranteed the intended level of accessibility, it may be prone to deadlocks when the user does not find the proper voice command to express his or her intent.

In [3], author explored the combined use of hand-tracking with voice input processed with automatic speech-recognition, to propose a multi-modal interaction experience in immersive VR. The devised system integrates four main functionalities. The *positioning*, which let the user to trigger a teleporting action by uttering phrases like "I'll go to X", being X a particular object or Point Of Interest (POI), the *object identification*, which allows user to use pointing to identify a particular POI without specifying the name ("I'll go there"), the *information mapping*, consisting in using the joint-input of pointing and voice to add custom labels to object within the VE, and finally the *disambiguation*, which happens when two or more objects fit the same physical verbal description (in that case, pointing can be used to identify the correct one).

Finally, in [4], a teleporting technique based on gaze direction and voice commands is used in the context of an approach for the creation of immersive Integrated Development Environments (IDEs) in VR. In particular the word "teleport" was used as trigger command, and the direction of the head was used as pointing action, obtaining a completely hands-free alternative to the pointing gesture.

Starting from these premises, two main categories of speech-based hands-free teleporting techniques can be identified: *speech-only* (voice commands or NLP to identify the POI), *direction-based* (voice commands as trigger), and *compound techniques*, combining the functionalities of the previous two. To the best of authors knowledge, a comparison of hands-free techniques belonging to the three categories for navigation of large VRTS has yet to be performed.

III. Materials and Methods

In this section, the implementation of the voice-based navigation system is described, along with the training scenario used for the experimental activity.

A. Speech Engine

For the speech recognition, Microsoft Speech 1.17.0[1] was used as the basis for the development of the capabilities required to support the speech-based navigation. In particular, the speech engine was designed to recognize and provide as output two elements: an *Intent*, representing the kind of operation that the user wants to perform, and an *Entity*, that is the object of the action. The engine supported 5 common intents in the context of industrial training (movement, select tool, deselect tool, open schematic and close schematic), however, for the purpose of the current evaluation, only the movement was made available inside the considered scenario. Intents and entities were managed through a Grammar Extensible Markup Language (GRXML) file which contained all the possible statements expressing the considered intents. The speech engine was interfaced with the VR application by means of a client-server approach based on WebSocket.

[1]Microsoft Speech: http://tiny.cc/clqquz

B. Case Study

The three techniques selected for the evaluation were Speech-only (S), Speech w/ Gaze (SG) and the Speech w/ Gaze & Descriptions (SGD), each provided with a specific disambiguation technique.

With S, as the name suggests, the user can only relies on voice input to navigate the VE:

- if the name of the POI is not ambiguous (e.g., "Take me to the yellow bin", and there is only one yellow bin in the scene), the user is teleported at the desired location;
- if the POI is specified in a generic way (e.g., "Take me to the bin", with more than one bin in the scene), the disambiguation logic manages the ambiguity;
- if the speech-recognition logic cannot recognize the requested action and/or the way the POI is specified, the error is signaled to the user.

The disambiguation logic for S consists in opening an User Interface (UI) element (a panel, shown in Figure 1a) displaying the various possibilities (e.g., the picture of every bin), and then asking to the user to resolve the ambiguity by specifying with voice commands the correct POI ("Which bin?").

To manage the disambiguation, when the VR application receives an *Intent-Entity* pair, the entity is processed by means of a dictionary which associates entities to objects inside the VE. Within the dictionary, a single element of the VE can be associated with an arbitrary number of entities, corresponding to the various synonyms which can be used referring to the same element. Along with specific entities, the dictionary also includes a number of generic entries, which will trigger the "Ambiguity" state. In particular, when the ambiguity is signaled, the UI panel is populated with the possible alternatives and then displayed to the user. At the same time, the ambiguity state is signaled to the speech engine, along with the indication of the generic entity, in order to set up the following voice recognition in the context of the disambiguation.

SG takes advantage of the user's head direction to cast a ray across the scene, which can be used to indicate a specific POI. Hence, the user can maintain the pointer on a specific object, and pronounce phrases like "Take me there". Similarly to S, if there are ambiguity in the POI selection (e.g., if the ray intersects more than one POI) the disambiguation is evoked. In this case, the user can interact with the UI panel with the same paradigm of the navigation technique, so by pointing with the head at the correct POI and saying expressions like "There". To facilitate the selection with the gaze, additional

graphical aids were included. In particular, occluding elements which are not among the possible POIs are automatically made transparent when hit with by the ray-cast, and a specific highlight is used to indicate the POI which is being selected, along with an arrow indicator positioned over the object itself.

Finally, SGD was obtained by combining the operation of the two previous techniques in a single approach. In particular, the user is free to use either natural language expressions or the gaze direction to identify the desired POI, and the system was designed trying to solve ambiguities by combining the two sources of information before launching the disambiguation technique (e.g., if the user is pointing the yellow bin, and says "Take me to the bin", the system will assume that it is the intended POI). In case of further ambiguities, the system will provide a disambiguation panel, and the user can interact with it with both approaches (S or SG).

C. Scenario

The test scenario was designed with two main objectives. To provide a fair test bench for all the techniques, avoiding situations where one of them cannot be used completely, and to stress the specific characteristics of all of them. As mentioned before, the VRTS represented a common training use case, that is an industrial hangar. The scenario, developed with Unity 2020.2 as a SteamVR application. The VE was populated with a number of virtual objects, some of them configured as possible POIs (reachable with the teleport), whereas the other were treated as context elements (e.g. obstacles).

The experience was organized as a set of atomic tasks, which order of exposition was randomized, each preceded by a preparation phase. The preparation phase consisted in transporting the user to a privileged POV (usually, a view from the top), from which are displayed the starting position (with a blue cirlce) and the POI that has to be reached to complete the given task. After that, the user is brought to the starting position, and the teleporting technique is enabled to let him or her to perform the task. At the completion, the user is again moved to the priviledged POV to begin the preparation for the following task. In should be noted that the user was not allowed to teleport to custom locations or wrong POIs. In this second case, the error is signaled (and logged), and after three consecutive errors the system provides a suggestion to panel to resolve the situation. The list of tasks with the relative description are depicted in Figure 1 and detailed in Table I.

IV. Experimental Methodology

The experiment was designed as a within-subjects user study involving 15 participants with ages ranging between 20 and 30 years, recruited among the staff of Leonardo S.p.A.. The hardware selected for the experiment was a Meta Quest 2 headset, used as a tethered OpenVR system by connecting it to a VR-ready laptop via the Oculus Air-Link capability.

At the beginning of the experiment, each participant was invited to fill in a demographic questionnaire, also intended to record their previous experience with VR applications and speech interfaces (e.g., voice assistants). After being instructed regarding the purpose of the evaluation, the participant was exposed to the three modalities following a latin-square order. Before each run, the following technique was described in detail, as well as the operation of the disambiguation logic.

As mentioned before, the experience is subdivided in various tasks which order was randomized for each run. After completing the VR experience with one technique, the participant is asked to fill a questionnaire composed by various sections. A first section, corresponding to the Subjective Assessment of Speech System Interfaces (SASSI) [17], is used to evaluate the usability of speech-based interfaces. A second section, corresponding to the System Usability Scale (SUS) questionnaire [18], investigates the overall usability of the VR system. The third section included custom questions regarding aspects which were not considered by the previous questionnaires (e.g. regarding the difficulty to reach POIs at given conditions). Finally, in the post-test section, participants have to rank the three techniques in order of preference, and possibly provide additional comments. The full questionnaire is available for download [2].

In addition to the subjective measures, the VRTS was designed in order to provide a number of objective measures in form of performance indicators. In particular, for each run, the application provided the average time per destination (from the start to the teleport), the number of errors in selecting the requested POI, and the number of commands not recognized by the speech-recognition algorithm (along with the total of these two types of errors).

V. Results

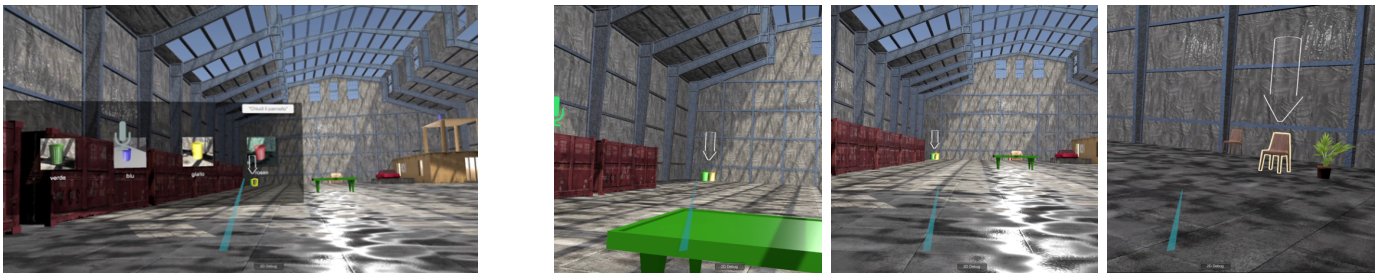
The results obtained for the subjective and objective metrics enounced in the previous section are discussed here below. One-Way repeated measures ANOVA and paired t-tests post-hoc with Bonferroni correction were applied to investigate the statistical differences.

A. Objective Metrics

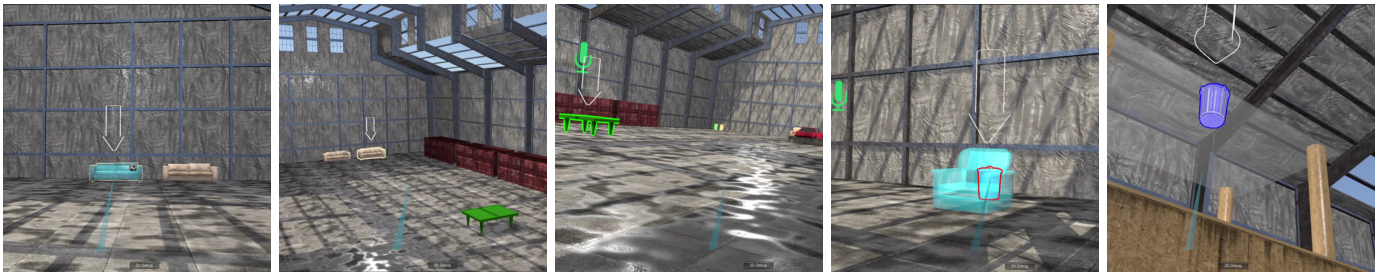
Results for the objective metrics are reported in Table II. Among the various measures considered in this study, the only one which resulted in a significant differences was the average number of errors related to the selection of a wrong POI, which were significantly lower in SG than with both S (1.62 vs 0.06, p-value $= .002$) and SGD (0.06 vs 1.31, p-value $= .003$).

In particular, with the SG technique, only 1 single wrong POI error was made in the whole group of 15 participants, whereas the other two techniques were characterized by a number of errors of the order of tens. This outcome is probably related to the fact that SG user can identify the teleportation target by pointing it with the head more clearly more univocally than with the other techniques, so it is less prone to mistakes in the selection of the POI. On the other hand, by using descriptions (i.e., with S and SGD), the user is not completely aware of the outcome of his or her expression.

[2]Questionnaire: http://tiny.cc/vgdhfcq



(a) Disambiguation UI (b) Task 1. (c) Task 2. (d) Task 3.



(e) Task 4. (f) Task 5. (g) Task 6. (h) Task 7. (i) Task 8.

Fig. 1. (a) UI panel shown to the user in order to manage the disambiguation; (b-h) Frames of the navigation tasks considered in the testing scenario.

Task #	Task	Condition	Target POI	Potential Wrong POI
1	Partially overlapped POIs	Medium Distance	Green bin at the corner	Yellow bin at the same corner
2	Partially overlapped POIs	Long Distance	Yellow bin at the corner	Green bin at the same corner
3	Similar POIs	Short Distance	Chair (closer, on the right, with the pot plant)	Chair (farther, on the left)
4	Similar POIs	Medium Distance	Couch (on the left, with the ball)	Couch (on the right)
5	Similar POIs	Long Distance	Couch (on the right)	Couch (on the left, with the ball)
6	Similar POIs	Different height	Table at the center	Table at the first floor
7	Visually occluded POI	Short Distance	Red bin	Any other bin
8	Visually occluded POI	Different height	Blue bin	Any other bin

TABLE I
DETAILS OF THE NAVIGATION TASKS CONSIDERED IN THE TESTING SCENARIO.

Objective Measure	**S**(SD)	**SG**(SD)	**SGD**(SD)	***p*-value**	S-SG	S-SGD	SG-SGD
Avg time per destination (s)	8.51(7.06)	7.47(5.16)	8.31(7.52)	.439	-	-	-
Avg errors (wrong POI)	1.62(1.80)	0.06(0.24)	1.31(1.45)	**.001**	**.002**	.595	**.003**
Avg errors (command not understood)	0.68(0.84)	1.00(1.27)	0.93(0.89)	.389	-	-	-
Avg errors (total)	2.31(2.19)	1.06(1.24)	2.25(1.64)	.082	-	-	-

TABLE II
AVERAGE RESULTS FOR THE OBJECTIVE MEASURES. STATISTICALLY SIGNIFICANT RESULTS ARE HIGHLIGHTED WITH A BOLD FONT.

1) Subjective Metrics: The results for the six sub-scales of the SASSI [17], expressed on a 7-points Likert scale (from strongly disagree to strongly agree), are reported in Figure 2.

As can be seen from the figure, the statistical analysis revealed significant differences for three out of six indicators.

Although no significant differences were found for the system response accuracy sub-scale, the interaction with the system in case of S was considered as significantly more efficient than SG (6.6 vs 5.73, p-value $= .006$).

Regarding the likeability, SG was perceived as significantly less likeable than S (6.29 vs 5.67, p-value $< .001$) and SGD (5.67 vs 6.23, p-value $= .018$). This outcome can be related to the fact that, although it allowed a higher accuracy, SG was based on a repetitive interaction scheme, which may be perceived as more and more tedious over time.

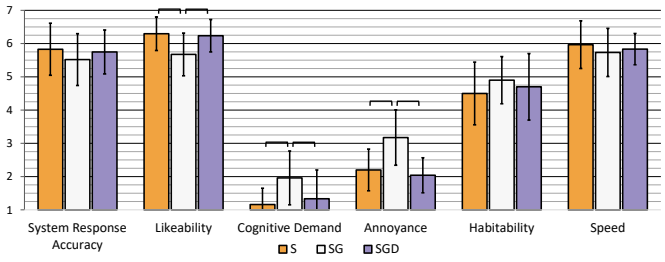


Fig. 2. Results for the SASSI [17]. Statistically significant results are marked with baffles, SD is expressed through bars.

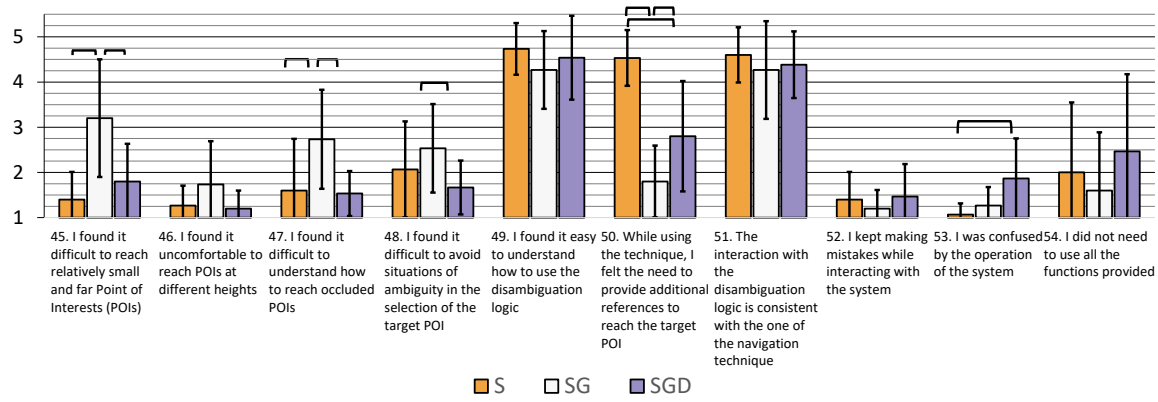


Fig. 3. Results of the custom section of the questionnaire. Statistically significant results are marked with baffles, SD is expressed through bars.

Interestingly, SG was considered as more cognitively demanding with respect to S (1.16 vs 1.96, p-value $< .001$) and SGD (1.96 vs 1.33, p-value $= .012$). Although this result may appear as counter-intuitive, it can be explained by looking at the questions belonging to this sub-scale. In fact, participants stated that they were calmer with S than with SG during the usage of the technique (6.33 vs 5.45, p-value $= .003$), and that SG required an higher level of concentration with respect to S (1.86 vs 3.46, p-value $< .001$) and SG (3.46 vs 2.33, p-value $= .001$). Although those statements were originally linked with cognitive load, participants interpreted them in terms of physical demand too. In particular, the higher level of concentration was probably perceived as a requirement to keep the gaze over the desired POV. Moreover, this action becomes increasingly difficult with the distance of the POI, and this increase of difficulty may lead to a state of agitation.

For what it concerns the annoyance, again SG performed significantly worse than S (2.2 vs 3.17, p-value $< .001$) and SGD (3.17 vs 2.04, p-value $< .001$).As this sub-scale concerns aspects such as boredom, frustration and inflexibility, this result could explain the the poor likeability of SG too.

Finally, no significant differences were observed for the habitability and speed sub-scales, although participants stated that they sometimes wondered if they were using the right word less frequently with SG than with S (4.46 vs 2.8, p-value $< .001$) and SGD (2.8 vs 3.13, p-value $< .001$). This outcome is not unexpected, as the vocabulary in case of SG was much more uniform and contained.

Regarding the SUS [18] section, in general, no significant differences were observed, as the three versions were characterized by fairly high total score values (between 80 and 90, greater than 71.1 threshold for *Good*).

Regarding the custom section of the questionnaire, expressed on a 5-point Likert scale from total disagreement to total agreement and reported in Figure 3, the significant differences provide a more insight about the previous outcomes. In particular, participants with SG found it more difficult to reach small and far POIs with respect to S (1.4 vs 3.2, p-value $< .001$) and SGD (3.2 vs 1.8, p-value $< .001$), probably due to the difficulty in keeping the gaze on tiny objects before uttering the phrase to trigger the teleporting. For the same reasons, SG made it more difficult to understand how to reach occluded POIs if compared to S (1.6 vs 2.73, p-value $= .007$) and SGD (2.73 vs 1.53, p-value $= .002$). As one could expect, the combination of functionalities of S and SG allowed SGD to mitigate the difficulty of avoiding ambiguities with respect to SG (2.53 vs 1.66, p-value $= .006$). In fact, with SG and SGD, the ambiguity of two or more POIs hit by the ray casted from the gaze can be easily solved by uttering a description of the target POI. Unsurprisingly, with SG, participants felt less the need to provide additional references to the desired POI with respect to both S (4.53 vs 1.8, p-value $< .001$) and SGD (1.8 vs 2.8, p-value $< .005$), but also with S with respect to SGD (4.53 vs 2.8, p-value $< .001$). This results suggest that SG, despite all its downsides, reduces the cognitive load related to the need of descriptions. Moreover, the SG functionality allows SGD to mitigate this issue, by providing an alternative way to solve ambiguous situations. However, participants also perceived the operation of SGD as more confusing than S (1.06 vs 1.86, p-value $= .005$), probably due to the fact that the combination of the functionalities of S and SG lead to a less uniform user experience.

Rank	S	SG	SGD
1st	40%	0%	60%
2nd	60%	20%	20%
3rd	0%	80%	20%

TABLE III
RANKING BY PREFERENCE: p-VALUE $= .002$, S-SG (p-VALUE $< .001$), S-SGD (p-VALUE $= .812$), SG-SGD (p-VALUE $< .001$)

Finally, regarding the ranking by preference (Table III), significant difference were observed between S and SG (p-value $< .001$), and between SG and SGD (p-value $= .001$), as S and SGD were usually chosen as first or second choice.

VI. Discussion and Conclusion

In this work, three speech-based hands-free navigation techniques for VR are deployed in a large indoor VE representing a

common industrial training scenario (i.e., a hangar), and compared in terms of performance by means of a within-subjects user study. The three techniques which the 15 participants were invited to test were a Speech-only (S) technique, based on the detection of utterances made by combinations of intents and entities (e.g., "Take me to the X"), in order to identify the Point Of Interest (POI) and trigger the teleport action, a Speech w/ Gaze variant (SG), in which the direction of the gaze is used to idetify the target POI, and the voice is used to trigger the teleporting (e.g., "Take me there"), and third technique obtained combining the functionalities previous two, labeled Speech w/ Gaze & Descriptions (SGD).

Results showed the undisputed superiority of S and SGD with respect to SG from many subjective point of views, from likeability, to cognitive demand, annoyance and preference. On the other hand, SG appeared to minimize the number of errors, in the form of number of selections of the wrong POI, and number of commands not understood by the speech engine. However, this small increase in terms of task performance comes at the cost of lower efficiency, pleasantness, enjoyability, control, calmness, as well as of higher repetitiveness, boredom, frustration and inflexibility. Interestingly, the compound technique (SGD) did not provide a significant benefit over S, except for a slightly reduced need for the provision of additional references to indicate the desired POI, but at the risk of causing more confusion due to the less uniform interaction scheme.

Future developments should be oriented towards overcoming the limitations of this work, thus by widening the evaluation to other more representative training use cases (e.g. with a larger number of less distinguishable elements), as well as to include in the experience other forms of navigation (e.g., teleporting to arbitrary 3D coordinates) to better challenge the performance of the considered techniques and the relative disambiguation logic. Furthermore, other speech-based techniques may be included in the comparison, for example considering also hands-busy implementations, such as the SG variants which are based on the pointing of the hand in place of the gaze, being them based on hand tracking or the use of a VR hand controller. Finally, it could be interesting to extend the the investigation to include also other relevant tasks other than navigation, such as the use of the speech to interact with objects (e.g., selection of tools, display and concealment of UI elements), the contemporary use of hands to perform manual tasks, as well as any other elements which may affect the performance of the considered techniques.

ACKNOWLEDGMENT

The authors want to acknowledge the help given by Leonardo S.p.A.[3] in the design and experimental phase, and of Emanuele Ceralli, who contributed to the development of the system and to the experimental phase.

[3]Leonardo S.p.A.: https://www.leonardo.com/

REFERENCES

[1] A. Garg, J. A. Fisher, W. Wang, and K. P. Singh, "Ares: An application of impossible spaces for natural locomotion in VR," in *Proc. 2017 CHI Conf. Ext. Abstr. on Human Factors in Computing Systems*, CHI EA '17, (New York, NY, USA), p. 218–221, Association for Computing Machinery, 2017.

[2] A. Ferracani, M. Faustino, G. X. Giannini, L. Landucci, and A. Del Bimbo, "Natural experiences in museums through virtual reality and voice commands," in *Proc. 25th ACM Int. Conf. on Multimedia*, MM '17, (New York, NY, USA), p. 1233–1234, Association for Computing Machinery, 2017.

[3] J. Sin and C. Munteanu, "Let's Go There: Voice and pointing together in VR," in *Proc. 22nd Int. Conf. on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2020)*, MobileHCI '20, (New York, NY, USA), Association for Computing Machinery, 2020.

[4] R. Mehra, V. S. Sharma, V. Kaulgud, S. Podder, and A. P. Burden, "Immersive IDE: Towards leveraging virtual reality for creating an immersive software development environment," p. 177–180, 2020.

[5] R. A. Bolt, ""Put-That-There": Voice and gesture at the graphics interface," in *Proc. 7th Ann. Conf. on Computer Graphics and Interactive Techniques (ACM SIGGRAPH)*, SIGGRAPH '80, (New York, NY, USA), p. 262–270, Association for Computing Machinery, 1980.

[6] R. Sharma, M. Zeller, V. Pavlovic, T. Huang, Z. Lo, S. Chu, Y. Zhao, J. Phillips, and K. Schulten, "Speech/gesture interface to a visual-computing environment," *IEEE Computer Graphics and Applications*, vol. 20, no. 2, pp. 29–37, 2000.

[7] "Rolls-Royce Opens BR725 Virtual Training Hangar." https://www.ainonline.com/aviation-news/business-aviation/2020-10-09/rolls-royce-opens-br725-virtual-training-hangar. Accessed: 2022-02-22.

[8] 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 *Proc. IEEE 9th Int. Conf. on Consumer Electronics (ICCE-Berlin)*, pp. 348–352, 2019.

[9] F. Corelli., E. Battegazzorre., F. Strada., A. Bottino., and G. P. Cimellaro., "Assessing the usability of different virtual reality systems for firefighter training," in *Proc. 15th Int. Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP,*, pp. 146–153, INSTICC, SciTePress, 2020.

[10] B. Sarupuri, S. Hoermann, M. C. Whitton, and R. W. Lindeman, "LUTE: A locomotion usability test environmentfor virtual reality," in *Proc. 10th Int. Conf. on Virtual Worlds and Games for Serious Applications (VS-Games)*, pp. 1–4, 2018.

[11] A. Cannavò, D. Calandra, F. G. Pratticò, V. Gatteschi, and F. Lamberti, "An evaluation testbed for locomotion in virtual reality," *IEEE Trans. on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 1871–1889, 2021.

[12] F. Buttussi and L. Chittaro, "Locomotion in place in virtual reality: A comparative evaluation of joystick, teleport, and leaning," *IEEE Trans. on Visualization and Computer Graphics*, vol. 27, no. 1, pp. 125–136, 2021.

[13] E. Bozgeyikli, A. Raij, S. Katkoori, and R. Dubey, "Point & teleport locomotion technique for virtual reality," in *Proc. 2016 Ann. Symp. on Computer-Human Interaction in Play*, CHI PLAY '16, (New York, NY, USA), p. 205–216, Association for Computing Machinery, 2016.

[14] F. G. Pratticò and F. Lamberti, "Towards the adoption of virtual reality training systems for the self-tuition of industrial robot operators: A case study at KUKA," *Computers in Industry*, vol. 129, p. 103446, 2021.

[15] F. G. Pratticò, D. Calandra, M. Piviotti, and F. Lamberti, "Assessing the user experience of consumer haptic devices for simulation-based virtual reality," in *Proc. 11th IEEE Int. Conf. on Consumer Electronics (ICCE-Berlin)*, pp. 1–6, 2021.

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

[17] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Language Engineering*, vol. 6, no. 3-4, p. 287–303, 2000.

[18] J. Brooke, "SUS: A 'quick and dirty' usability scale," *Usability Evaluation in Industry*, p. 189, 1996.