"Paint that object yellow": Multimodal Interaction to Enhance Creativity During Design Tasks in VR

Erik Wolf
University of Würzburg
Human-Computer Interaction
Würzburg, Germany
erik.wolf@uni-wuerzburg.de

Sara Klüber
University of Würzburg
Psychological Ergonomics
Würzburg, Germany
sara.klueber@uni-wuerzburg.de

Chris Zimmerer
University of Würzburg
Human-Computer Interaction
Würzburg, Germany
chris.zimmerer@uni-wuerzburg.de

Jean-Luc Lugrin
University of Würzburg
Human-Computer Interaction
Würzburg, Germany
jean-luc.lugrin@uni-wuerzburg.de

Marc Erich Latoschik
University of Würzburg
Human-Computer Interaction
Würzburg, Germany
marc.latoschik@uni-wuerzburg.de

ABSTRACT

Virtual Reality (VR) has always been considered a promising medium to support designers with alternative work environments. Still, graphical user interfaces are prone to induce attention shifts between the user interface and the manipulated target objects which hampers the creative process. This work proposes a speech-and-gesture-based interaction paradigm for creative tasks in VR. We developed a multimodal toolbox (MTB) for VR-based design applications and compared it to a typical unimodal menu-based toolbox (UTB). The comparison uses a design-oriented use-case and measures flow, usability, and presence as relevant characteristics for a VR-based design process. The multimodal approach (1) led to a lower perceived task duration and a higher reported feeling of flow. It (2) provided a higher intuitive use and a lower mental workload while not being slower than an UTB. Finally, it (3) generated a higher feeling of presence. Overall, our results confirm significant advantages of the proposed multimodal interaction paradigm and the developed MTB for important characteristics of design processes in VR.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **Interaction techniques**; *Empirical studies in HCI*.

KEYWORDS

Speech and Gesture, Creativity, Design, 3D User Interfaces

ICMI '19, October 14–18, 2019, Suzhou, China

ACM Reference Format:
Erik Wolf, Sara Klüber, Chris Zimmerer, Jean-Luc Lugrin, and Marc Erich Latoschik. 2019. "Paint that object yellow": Multimodal Interaction to Enhance Creativity During Design Tasks in VR. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353724

1 INTRODUCTION

In recent years, designers started to use VR applications like Enscape [19] or IrisVR [23] to design products. VR inherently supports interactive workflows as well as interactive and immersive product presentations. Typical VR display systems, e.g., large stereoscopic displays like CAVEs [8], or head-mounted displays (HMDs) [46], support head-tracking. This allows users to change their view perspective freely to inspect a virtual scene and target objects from any position they want within the limits of the given display by just moving their head as they would do in the physical world. This close resemblance between physical movement and the generated user-centered visual feedback of the virtual scene creates a strong spatial awareness and presence [20]. The increased spatial awareness and presence, in turn, foster a holistic comprehension of the designed product [39, 43].

As a result, the visual appearance and impression of virtual objects in VR design applications resemble the visual impression of real objects in the physical world much closer compared to desktop systems. Still, VR retains many advantages of digitized workflows regarding efficiency and effectivity, such as lower costs, enhanced configuration options, and dynamic simulations [50]. For example, IrisVR helped to identify and fix design inconsistencies during the design of water recycling centers [24].

Overall, the intuitive use of the user interface is an important factor for a successful design process as it tends to promote productivity, creativity, and enjoyment. Compared

to traditional desktop software, VR applications already provide a more natural way of viewing and inspecting a product which closely resembles the way people view and inspect objects in the physical world. However, visual feedback and inspection are only one aspect of the required interactions of computer-supported design applications. Obviously, users also will have to change and manipulate the objects during the design process. The impact of creativity support tools [44], and the productivity of applications on the design process have been examined. In the meantime, the effects of 3D interaction techniques on creativity have received less attention despite their importance during the creative process of product design. This work explores the potential of a multimodal interaction metaphor that combines speech and gesture input within an immersive VR environment regarding its creativity-enhancing capabilities.

2 RELATED WORK

A factor that positively correlates with creativity is the feeling of flow [53]. Flow has an indirect effect on creativity through exploratory behavior and positive affect [53] and is often reported by people engaged in creative tasks [45]. As described by Csikszentmihalyi and Sawyer [9], flow reflects attentional processes and is achieved when attention is fully invested in the task at hand, and the task difficulty matches the abilities. Subsequently, a flow-enhancing environment should support users to focus on the task without interruptions while not overwhelming them regarding the difficulty of the task. To support this, applications, or more precisely, interaction techniques with a high degree of usability are required [40]. Effectivity, efficiency, and user satisfaction are key aspects of usability [25]. Intuitive usage is considered one strong promotor for these aspects [22]. An interaction technique may be intuitive, especially if it builds on subconscious knowledge resulting in lower mental workload. A good example of this is the head-tracked view manipulation of typical VR systems based on physical head movements learned in the real world. Additionally, the usability factor of efficiency also has an impact on productivity.

Another factor associated with flow is the induced presence [20], where a high degree of induced presence has contributed to a higher degree of flow [26, 27]. Overall, an interaction technique used for design processes should therefore increase flow, usability, and presence in order to promote creativity-enhancing conditions and to allow users to fully focus on a given task and the related object(s) of interest.

Commonly used control techniques for VR applications are menu systems with one degree of freedom adapted from 2D desktop environments, often virtually attached to a controller device [11] or stationary to the virtual camera within the user's field of view. However, menus force users to shift their visual attention and concentration to the menu for searching the correct menu item rather than to keep it on the environment or the object of interest [41], potentially breaking the flow of creativity. This problem has been known for 2D desktop systems and is considered aggravated for VR. VR is characterized by a dynamic and extended user-centered visual field and typically exploits a simulation of visual depth based on stereoscopy. Current stereoscopy methods are prone to the vergence-accommodation conflict [30], which will additionally worsen the negative impact of a visual attention shift. These drawbacks concern every menu technique, even radial menus [7, 17] which otherwise have been shown to be superior in terms of efficiency and usability compared to traditional menus.

An interaction technique which allows staying focused on the object of interest is multimodal interaction as pioneered by Bolt's 'Put-That-There' [3]. Multimodal interaction is the combination of different modalities, which can result in usability advantages, such as increased efficiency due to reduced mental workload [35, 42], as well as increased user satisfaction [12, 37], compared to unimodal interaction. Despite the research on multimodal interaction conducted since Bolt's pioneering work, there is still a need for further research in different directions, e.g., reliable multimodal processing systems and usable multimodal applications [13]. A comparable approach of an instruction-based speech and gesture interface has been investigated, for example, in the area of self-driving cars [49]. However, comparisons between multimodal speech and gesture interfaces and unimodal menu-based interfaces in VR are still sparse and to the best of our knowledge do not exist in the context of current state-of-the-art design applications.

Given the potentially higher usability and the absence of required visual attention shifts as present in menu techniques, a multimodal interface promises to support a higher feeling of flow. Multimodal interfaces center around users' natural behavior and communication capabilities [38]. As such, they may build on subconscious knowledge more than a menu interaction does, further suggesting a reduced mental workload. Additionally, the natural interaction in the virtual environment may increase the feeling of presence. Multimodal speech and gesture interfaces for VR have continuously been explored in the past, e.g., [2, 31–34], but are still considered a niche interaction metaphor in comparison to menu-driven interactions.

Approach and Hypothesis

Recent expert interviews identified potential benefits of unimodal voice shortcuts for desktop-based design application and highlighted promising possibilities for designers' workflows [29]. We extend this idea and combine the proposed

benefits of multimodal input with the intuitive visual feedback as provided by immersive VR to exploit potential benefits of the multimodal interaction paradigm for VR design applications. Our approach uses a typical VR visualization and inspection method based on a head-tracked HMD with a concurrent speech and gesture input which provides instruction-based input capabilities inspired by Bolt's work. In combination, VR and multimodality should promote flow, usability, and presence, and hence constitute appropriate interaction techniques that produce creativity-enhancing conditions and thus seamlessly let a user control creativity-demanding design applications.

We evaluate the multimodal interaction approach in comparison to a typical menu-oriented interface technique. The experimental evaluation compares the two techniques in terms of flow, usability, and presence as our main target factors to explore the potential benefits of the two interface metaphors for creativity-demanding design applications. We expect the MTB to better support flow, usability, and presence than the UTB. Table 1 illustrates the detailed variables and hypothesis of the evaluation. The table reports the derived results as a look-ahead for a quick overview and orientation. The MTB turned-out to be superior in comparison to the UTB in all aspects while it did not hamper the interaction time and hence could be considered equally efficient.

Table 1: Overview of variables, hypotheses and results.

Variable	Hypotheses	p
Flow		
(H1) Flow survey score	UTB < MTB	.005*
(H2) Relative subjective duration	UTB > MTB	.004*
Usability		
(H3) Intuitive use survey score	UTB < MTB	.014*
(H4) Workload survey score	UTB > MTB	.007*
(H5) Object modification time	UTB > MTB	.364
Presence		
(H6) Presence survey score	UTB < MTB	.007*

Asterisks indicate significance.

3 SYSTEM DESCRIPTION

Task and Use Case

The experimental task is based on the use-case of an object modification process as present in current design applications. The goal is to manipulate the visual appearance of one object (Figure 1, right) to match the appearance of a provided sample object (Figure 1, left). The task is inspired by the work of Bowman and Wingrave [6] and is designed to avoid dependence on participants' individual creativity deliberately. Size, color, and texture of objects can be changed to five possible values each. Either one, two, or all three properties have to be changed to match the sample object's appearance. The virtual environment consists of a simple room that enables realistic object shape, color, and depth perception. The objects are placed on two pedestals while the participant stands on a third pedestal. In order to perform an object manipulation, the participant either has to point to the respective object during the manipulation, i.e., highlighting it, or select it in advance. Both, the UTB and the MTB use a ray-casting technique for pointing [35]. Hands and controllers are shown in VR, and a soft vibration signals the intersection of the pointing ray with an object. A frame around the object provides visual feedback about whether the object is selected (purple bold), highlighted (yellow), or selected and highlighted (yellow bold). The application provides acoustic

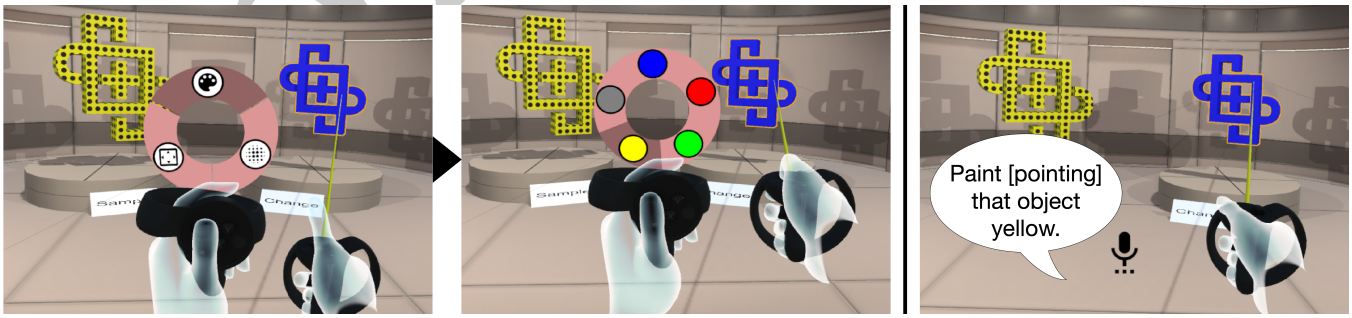

Figure 1: Experiment overview: Participants were asked to change the size, color, and texture of the right object to match the left object using a unimodal (radial menu) or multimodal (speech and gesture) toolbox. The two images on the left show the two-step process of changing the color of an object using the UTB. First, the action 'change color' must be selected from the radial menu (left). Second, the respective color, in this case yellow, must be selected while pointing on the object (middle). The image on the right depicts the same action using the MTB. The respective action can be invoked by the multimodal utterance: *'Paint [pointing] that object yellow.'*

feedback for each object change and match. For both, UTB and MTB, a help menu was available to provide information about the use of the toolboxes.

Implementation

The application's architecture uses two subsystems dedicated for input processing and VR simulation which are interconnected by a transport layer. The overall architecture is illustrated in Figure 3. The virtual environment (Physics, Renderer), the application logic (Logic), and the UTB are realized with *Unity 2017.4.8f1* [48], a VR-capable game and 3D development platform. An *Occulus Rift* HMD is used to present the virtual environment, and two *Occulus Touch* controllers for both, to (1) operate the UTB and to (2) provide gesture input for the MTB. The open-source platform *Simulator X* [33] provides the necessary extensions to implement the Multimodal System (MMS) for the MTB. Simulator X's implementation of a *concurrent Augmented Transition Network* (cATN) [55] is used to define possible multimodal utterances and conduct multimodal fusion. It is a successor of the *temporal Augmented Transition Network* (tATN) [32] which is tailored for semantic grounding and concurrent processing, e.g., the concurrent analysis of speech and gesture input. The cATN is openly available for research [56]. Machine learning approaches exhibit great potential in early fusion tasks, e.g., in computer vision, natural language processing, robotics, and information retrieval [15]. For late fusion, descriptive fusion methods are still predominant, however, they are open for supplementation with machine learning [55]. Drawbacks of descriptive fusion methods include that the developer has to define all possible utterances the system can recognize manually and that descriptive fusion methods are subject to issues regarding computational complexity [28]. We chose this procedural fusion method since it supports the rapid development of multimodal interfaces without relying on time-costly training and optimization as common in machine learning-based solutions.

The *Microsoft Speech SDK* is used for the required speech recognition and speech input. Recognition results are transmitted to the multimodal integration, i.e., the processing via the cATN, using the *VRPN* protocol [47]. Gesture input is recognized by a dedicated subsystem (Gesture) implemented in Unity and is represented in the application state, i.e., in a `User` game object. This information, alongside other context information necessary for multimodal fusion, e.g., a game object's size, color, texture, and whether it is selected, is bidirectionally synchronized with Simulator X by means of a TCP transport layer as proposed by Wiebusch et al. [52]. This interaction context can be polled by the cATN on demand instead of being pushed to the cATN in large numbers. All systems run on a state-of-the-art VR-capable PC.

Unimodal Toolbox

We use a two-level radial menu implementation provided by the *Virtual Reality Toolkit V3.2.0* [51] for the UTB. The first menu level (Figure 1, left), displays the possible modification properties, the second its possible values (Figure 1, middle). Figure 2 provides an overview of all menu items. The user indicates an object to modify by pointing at it with a ray attached to the right controller. Pressing the `A` button while pointing on an object (automatically highlighted) will confirm the object selection. The menu is bound to the left controller's position and can be opened by pressing its thumbstick. A menu item can be chosen by using the thumbstick, where a visual marker indicates which menu item is currently selected. The left controller's trigger button serves as the *back* and *close* menu function. The menu automatically closes after selecting an property (e,g., color).

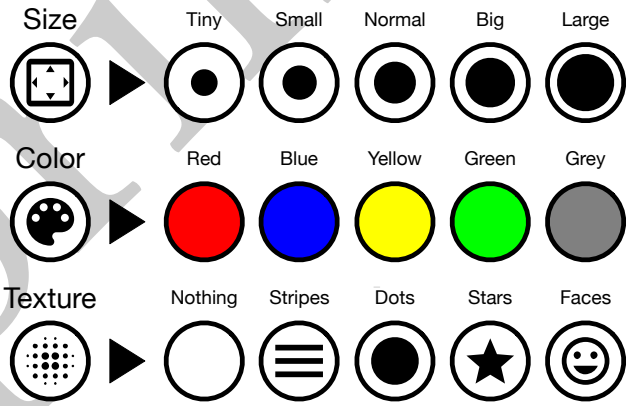

Figure 2: Unimodal Toolbox (UTB) commands using a radial menu's proposing actions (left) and properties (right).

Multimodal Toolbox

The MTB (Figure 1, right) consists of two modalities, speech and gesture, which can be used synergistically. Table 2 provides an overview of all supported multimodal utterances for changing the size, color, and texture of objects, as well as selecting them. If an object is selected, it can be referenced with *'it'*. Otherwise, the participant has to point to an object while simultaneously saying *'that object'* or simply *'that'*. The speech recognition indicates active processing through a pulsing microphone icon to give participants feedback and indicate when they can interact. Multimodal utterances are defined by means of the cATN's description language which automatically generates respective transition graphs for the cATN parser. Figure 4 provides a visual representation of the two transition graphs used to parse the multimodal utterances for changing an object's size, color, and texture. A transition graph usually consists of states, e.g., `S1`, that are

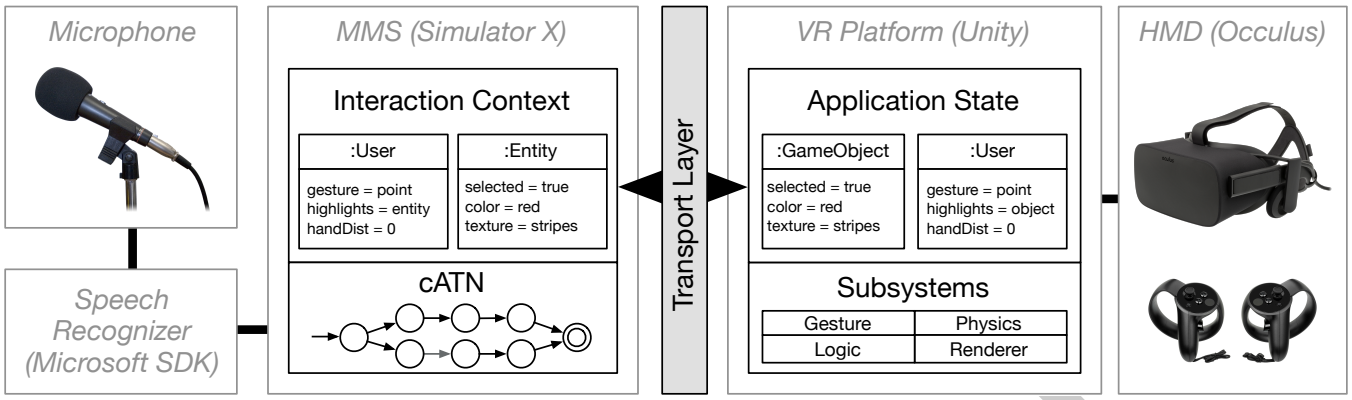

Figure 3: The application's architecture is a platform-compound, consisting of the multimodal system *Simulator X* and the VR platform *Unity*. The virtual environment, the application logic, as well as the UTB are implemented with Unity. The MTB relies on Simulator X's of a *concurrent Augmented Transition Network* to perform multimodal fusion. The *Microsoft SDK* provides speech input, while performed gestures are polled by the cATN from the with Unity synchronized interaction context.

Table 2: MTB commands using speech and *gestures*: *[pointing]* and *[size]* (two-handed).

Size	Utterances	Make <Object> <Property> | Make <Object> this *[size]* size.
	Object	that *[pointing]* object | that *[pointing]* | it
	Property	tiny | small | normal | big | large | small
Color	Utterances	Paint <Object> <Property>.
	Object	that *[pointing]* object | that *[pointing]* | it
	Property	red | blue | yellow | green | gray
Texture	Utterances	Texture <Object> with <Property>.
	Object	that *[pointing]* object | that *[pointing]* | it
	Property	nothing | stripes | dots | stars | faces
Selection	Utterances	Select <Object>.
	Object	that *[pointing]* object | that *[pointing]*

connected by arcs, e.g., `Verb`. Each arc consists of a condition and a function. The cATN uses potentially multiple *concurrent cursor* to represent active states in the transition graph. A multimodal utterance has been successfully recognized if a cursor reaches an `End` state. In response to an input, a cursor is able to transition from one state to another, if the bridging arc's condition is satisfied. In this case, the arc's function is executed. Each cursor contains a set of registers in which both, parsed input as well as interaction context information, can be stored. The cATN uses semantics-based state- and behavior-management techniques [16] to realize an access scheme to the interaction context. The condition of an arc can check the input or directly retrieve information about the application state from the interaction context, e.g., where the participant points, and stores this information in the cursor. In addition, an arc's function is able to manipulate the interaction context, e.g., changing the color of an object.

Figure 4 illustrates this process for the command: *'Paint [pointing] that object yellow.'* First, the parser checks if the input *'Paint'* satisfies the arc `Verb`'s condition, i.e., if the speech token of word type *verb*. The cursor then transitions to the state `S1` and stores the speech token in one of its registers. `Obj` is a dedicated sub-arc. In order to transition from `S1` to `S2` over `Obj`, the cursor has to transition through the lower transition graph, i.e., has to reach the `EndObj` state. The first part of the lower transition graph (from `Obj` to `S3`) depicts the network abstraction construct *Split-Merge* of the cATN. A cursor can only traverse from `Obj` to `S3` if both a speech token *'that'* as well as a pointing gesture are processed while satisfying a predefined temporal condition, e.g., both inputs have to be occurred within a 500 ms interval. After a successful *Split-Merge* an optional speech token *'object'* may follow. The `resolveObj` arc is a special arc that

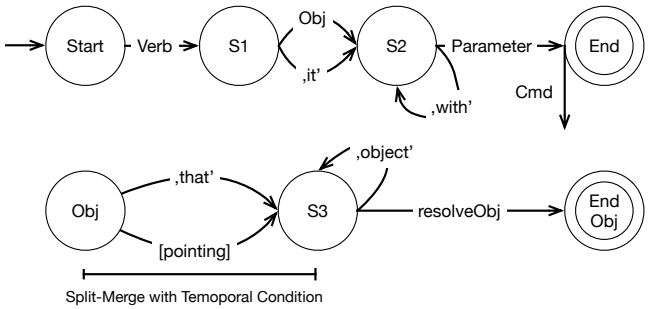

Figure 4: Two simplified cATN transition graphs for recognizing the multimodal utterances for changing an object's property. The upper graph defines the structure of the utterance. The lower graph is dedicated for recognizing noun phrases accompanied by a pointing gesture. It is utilized in the upper graph between state `S1` and `S2`.

is not triggered by new input, but instead the arrival of a cursor on its originating state. This arc performs semantic integration with its function by accessing the interaction context, retrieving the object to which the participant refers, and storing its reference in the cursor's register. Finally, the parser checks for a property, i.a., *'yellow'*. The cursor reaches the `End` state and triggers the dedicated feedback arc `Cmd`. Its condition checks the contents of the cursor's register and performs the necessary changes in the interaction context, i.e., changing the objects color property to the value yellow. This change in the interaction context is then synchronized to the VR platform where the virtual environment is changed accordingly.

4 STUDY

The experimental design was within-subjects and the independent variable was the type of toolbox with two levels: UTB and MTB. The dependent variables are divided into three categories: Flow, usability, and presence.

Flow. We used a survey on the two flow characteristics enjoyment and concentration based on the work of Ghani and Deshpande [18]. This provides an indication of whether the interaction techniques supported the feeling of flow by promoting object concentration and high enjoyment. Both characteristics were captured by four questions, each of which used a 5-point Likert scale with a range from 1 to 5 (5 = high feeling of flow). Since high levels of flow give the impression that time passes more quickly, we also calculated the relative subjective duration (RSD) [10] the experimental task took. RSD gives more insights on whether participants experienced the feeling of flow by calculating the percentage difference between perceived task duration and actually measured task duration. The lower the percentage, the higher the flow.

Usability. To measure whether the interaction technique was intuitive and therefore characterized by a high degree of usability, we used the *Questionnaire for Subjective Consequences of Intuitive Use* (QUESI) [36]. The QUESI contains 14 items that range from 1 to 5 (5 = high intuitiveness) structured in 5 dimensions: mental workload, the achievement of goals, the perceived effort of learning, familiarity, and the perceived error rate. We measured mental workload using the SEA scale [14], a German version of the Rating Scale Mental Effort [1, 54], to investigate intuitive use further. SEA is a single item scale ranging from 0 to 220 (220 = high workload). Lastly, we also measured a usability factor more related to productivity, namely efficiency, by recording the duration that a participant needed for each modification task. As a modification trial consisted of either one, two, or three modifications, we divided the total trial duration by the number of modifications.

Presence. In order to measure perceived induced presence, we used the presence score by Bouchard et al. [4, 5]. This score measures presence on a single item scale ranging from 0 to 10 (10 = high presence). The SEA and the presence score were measured mid-immersion multiple times.

Procedure

Participants followed a one-hour experimental procedure illustrated in Figure 5. After participants read information about the experiment, gave consent, and filled in a demographic survey, they tested both toolboxes in a counterbalanced test procedure shown in Figure 5, right. In this procedure, participants got familiar with the toolboxes by watching an explanation video on a separate screen and performing a two-minute free training phase in the virtual environment in which they got familiar with the respective toolbox. In a second training phase, they familiarized with the experimental task by performing ten modification trials. Only in the training phases, participants were allowed to ask questions about the interfaces. The test phase consisted of 10 x 3 counterbalanced modification trials. The mid-immersion mental workload score was captured after trial 5, 10, and 15 and the mid-immersion presence score after trial 10, 20, and 30. At the end of each toolbox test, but before participants took off the HMD, they were immediately asked to indicate the perceived duration of the experiment before they could check the time. This measurement was used in combination with the actual task duration to calculate the RSD. After participants left immersion, they finished the first toolbox test by filling in the *Flow* and *Usability* surveys. Then, participants continued with the second toolbox test. In the end, participants filled in a post-survey on what toolbox they preferred, with what toolbox they could better concentrate on the object and the task, and an open question for comments regarding the toolboxes. Comments from the open question are reported in the qualitative observations section.

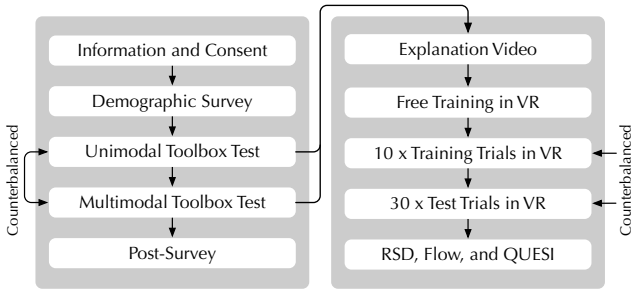

Figure 5: The general experimental procedure and the toolbox test procedure. Toolboxes were tested in a counterbalanced within-subjects design and trials were counterbalanced between one, two, and three required modifications.

5 RESULTS

In total, 33 university students participated in the experiment conducted in a laboratory at the University of Würzburg and received course credit in return. Two participants were excluded due to technical issues and another one due to the feeling of malaise. The 30 remaining participants (13 male, 17 female) were aged between 18 and 29 (M = 21.2), all German native speakers, and had no hearing or vision impairments. Two participants experienced VR for the first time, 24 one to ten times, and four used VR more than ten times.

Quantitative Measurements

Captured data met parametric testing requirements and hypothesis were directed. Therefore, we calculated one-sided paired-sample t-tests. We adjusted the alpha-level for each test according to the Bonferroni-Holm test procedure [21].

Flow. H1 has been confirmed as the flow survey showed a significantly higher score for the perceived feeling of flow for the MTB (M = 3.95, SD = 0.62) compared to the UTB (M = 3.65, SD = 0.8), $t(29)$ = 2.791, p = .005, dz = .51 (Figure 6, left). The alpha-level was α = 0.01. In line with H2, participants perceived the duration of the toolbox test (in relation to the actual time needed) significantly shorter when using the MTB (M = 12.97 %, SD = 41.79 %) than when using the UTB (M = 32.07 %, SD = 47.92 %), $t(29)$ = 2.87, p = .004, dz = .516 (Figure 6, right). The alpha-level was α = 0.0083.

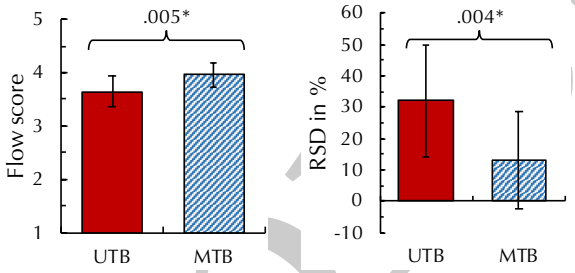

Figure 6: The left chart shows the average flow survey score, the right one the average RSD. Error bars represent 95 % confidence intervals.

Usability. Confirming H3, participants rated the MTB (M = 3.77, SD = 0.67) significantly better in intuitive use (QUESI) than the UTB (M = 3.38, SD = 0.71), $t(29)$ = 2.31, p = .014, dz = .428 (Figure 7, left). The alpha-level was α = 0.025. H4 has been satisfied since participants using the MTB (M = 52.63, SD = 25.99) perceived a significantly lower mental workload (SEA) than when using the UTB (M = 72.27, SD = 40.66), $t(29)$ = 2.64, p = .007, dz = .495 (Figure 7, right). The alpha-level was α = 0.0125. H5 has been rejected as the MTB (M = 5.398 s, SD = 0.891 s) did not allow for a significantly faster object modification than the UTB (M = 5.322 s, SD = 1.234 s), $t(29)$ = 0.35, p = .364, dz = .064 (Figure 8, left). The alpha-level was α = 0.05.

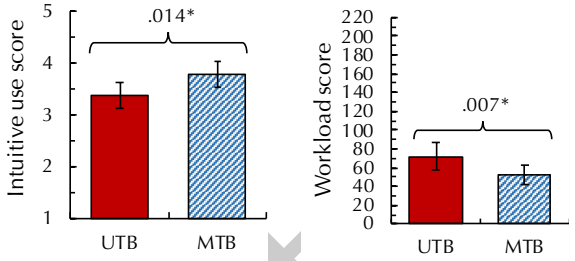

Figure 7: The left chart shows the average intuitive use (QUESI) score, the right one the average mid-immersion mental workload (SEA) score. Error bars represent 95 % confidence intervals.

Presence. According to H6, participants reported a significantly higher perceived score in the presence survey when using the MTB (M = 7.4, SD = 1.68) than when using the UTB (M = 6.8, SD = 2.02), $t(29)$ = 2.61, p = .007, dz = .317, (Figure 8, right). The alpha-level was α = 0.0167.

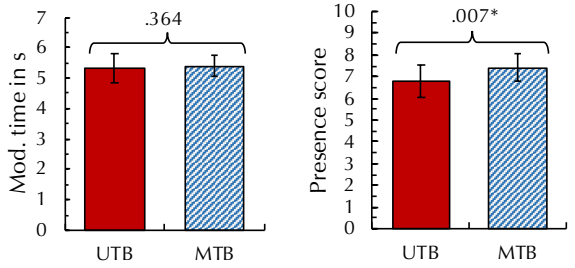

Figure 8: The left chart shows the average time per modification, the right one the average mid-immersion presence score. Error bars represent 95 % confidence intervals.

Qualitative Observations

During the experiment, various observations and comments were made by the experimenter and the participants. The experimenters had the feeling that particularly extroverted participants, who talked often and tended to comment on the experiment, preferred MTB over UTB. Participants who were insecure about VR and the technology used seemed to find it much easier to use the MTB. Conversely, people who preferred the UTB reported that they are more used to menus than speech interactions. Participants mentioned several times that they found MTB to be more intuitive and less distracting and could therefore concentrate better on objects. Most were surprised that the MTB worked so smoothly. However, they criticized the limited vocabulary and grammar. Some participants suggested combining the two toolboxes.

6 DISCUSSION

We compared creativity influencing flow characteristics, usability factors, and perceived presence of a UTB and a MTB in a design-inspired modification task realized in an immersive virtual environment. As expected, the flow measurements showed a significantly higher score in the flow survey (H1) and a lower relative subjective task duration (H2) in favor of the MTB. For the usability factors, we confirmed our hypotheses that the MTB allows for a significantly higher intuitive use (H3) than the UTB. Also, the mental workload felt significantly lower (H4) when using the MTB while object modification times were, contrary to our hypothesis (H5), not different between both toolboxes. We also confirmed the expected significant higher presence score (H6) for the MTB compared to the UTB. Quantitative results were supported by the users' statements as a majority of 20 participants preferred MTB over UTB. A significant majority of 25 participants stated that they could better concentrate on the objects when using the MTB. Replies to the final open question included statements like: '*I preferred the MTB. The UTB was potentially faster, however, the MTB did suit the task much better, and it was way more intuitive.*'

When using the MTB, users did not have to switch their visual attention between modification object and menu and did not have to search within the menu. By avoiding the mentally demanding and time-costly context switch and the following search task, users directly reported a lower mental workload. Participants could stay focused on the object which paid-off in a higher concentration towards the objects. The high intuitive use of the MTB allowed for a seamless control of the system which kept the concentration on the task. This supported a high level of enjoyment and usability, ultimately leading to a higher feeling of flow, which was supported by the RSD and flow survey measurement. The low RSD is also an indication that participants perceived the task as more difficult when using the UTB [10], which again is in line with the mental workload measurement.

The QUESI results indicated a lower subjective mental workload, a higher perceived achievement of goals, a lower perceived effort of learning, and a higher familiarity in favor of the MTB. For the fifth dimension, participants perceived a higher error rate for the MTB compared to the UTB. This can be mostly attributed to the probabilistic output of the speech recognizer which sometimes misinterpreted similar-sounding properties (e.g., *green* as *gray*). Therefore, participants had to repeat the sentence which is also an explanation for the not significantly lower modification time for MTB. However, the progressive development of speech SDKs will potentially lead to a further improved use of MTBs by reducing the error rate and consequently, the average object modification time.

Altogether, the results indicate that multimodal interfaces allow for a higher usability, a more profound flow experience, and a higher perceived induced presence while not being slower than the menu-based interface. Given the new perspective of design and creativity rather than productivity, these results are a first indication that multimodal interfaces may be more appropriate for design use cases by fostering creativity and enjoyment, allowing users to concentrate on their design process rather than the system interaction.

Limitations and Future Work

Our results confirm the benefits of the multimodal approach. Nonetheless, this initial exploration has the following limitations that open up space for future research:

(1) The experimental task deliberately avoided asking participants to design objects completely or to be creative to not depend on individuals' creativity. Our future experiments will include tasks which demand creativity, with a broader set of tools and commands to get even closer to current VR design applications.

(2) We did not find differences in *object modification time*. This could be caused by the small number of available modification properties and values. Future experiments should therefore compare modification times in relation to the number of properties and values.

(3) Multiple individual factors can contribute to a unimodal or multimodal toolbox preference. Further experiments should explore factors like the user's personality, experience with a system, or learnability to identify further use cases and required adaptions.

7 CONCLUSION

In this work, we proposed a multimodal interaction paradigm based on the combination of speech and gesture input within immersive VR to enhance creative tasks for interactive design applications. We motivated the approach derived from various findings in the related work to identify the three factors flow, usability (intuitiveness, mental workload, and task completion time), and presence as fostering creativity-enhancing conditions. Consequently, these factors served as target factors to assess the overall quality of our approach.

Altogether, our results confirm significant advantages of the proposed multimodal interaction paradigm and the developed MTB for important characteristics of design processes in VR. The results provide guidance for system architects and inspire developers of design applications by giving insights into promising alternative interfaces. Our future work will replicate this experiment with a toolbox that proposes a larger set of modifications, and with a non-constrained creative task where participants can freely design a new product without replicating an existing one.

REFERENCES

[1] Albert G. Arnold. 1999. Mental effort and evaluation of user-interfaces: a questionnaire approach. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I-Volume I*. L. Erlbaum Associates Inc., 1003–1007.

[2] Mark Billinghurst, Jesus Savage, Paul Oppenheimer, and Chuck Edmond. 1996. The Expert Surgical Assistant: An Intelligent Virtual Environment with Multimodal Input. In *Medicine Meets Virtual Reality IV: Health Care in the Information Age*. IOS Press, Amsterdam, 590–607.

[3] Richard A. Bolt. 1980. "Put-that-there": Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3 (July 1980), 262–270. https://doi.org/10.1145/965105.807503

[4] Stéphane Bouchard, Geneviéve Robillard, Julie St-Jacques, Stéphanie Dumoulin, Marie-Josée Patry, and Patrice Renaud. 2004. Reliability and validity of a single-item measure of presence in VR. In *Haptic, Audio and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on*. IEEE, 59–61.

[5] Stéphane Bouchard, Julie St-Jacques, Geneviève Robillard, and Patrice Renaud. 2008. Anxiety increases the feeling of presence in virtual reality. *Presence: Teleoperators and Virtual Environments* 17, 4 (2008), 376–391.

[6] Doug A. Bowman and Chadwick A. Wingrave. 2001. Design and evaluation of menu systems for immersive virtual environments. In *Virtual Reality, 2001. Proceedings. IEEE*. IEEE, 149–156.

[7] Jack Callahan, Don Hopkins, Mark Weiser, and Ben Shneiderman. 1988. An empirical comparison of pie vs. linear menus. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 95–100.

[8] Carolina Cruz-Neira, Daniel J. Sandin, Thomas A. DeFanti, Robert V. Kenyon, and John C. Hart. 1992. The CAVE: audio visual experience automatic virtual environment. *Commun. ACM* 35, 6 (1992), 64–72. https://doi.org/10.1145/129888.129892

[9] Mihaly Csikszentmihalyi and Keith Sawyer. 2014. *Shifting the Focus from Individual to Organizational Creativity*. Springer Netherlands, Dordrecht, 67–71. https://doi.org/10.1007/978-94-017-9085-7_6

[10] Mary Czerwinski, Eric Horvitz, and Edward Cutrell. Year. Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 conference*, Vol. 2. 167–170.

[11] Raimund Dachselt and Anett Hübner. 2007. Three-dimensional menus: A survey and taxonomy. *Computers & Graphics* 31, 1 (2007), 53–65.

[12] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. *Multimodal interfaces: A survey of principles, models and frameworks*. Springer, 3–26.

[13] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. Multimodal interfaces: A survey of principles, models and frameworks. In *Human machine interaction*. Springer, 3–26.

[14] Karin Eilers, Friedhelm Nachreiner, and Kerstin Hänecke. 1986. Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. *Zeitschrift für Arbeitswissenschaft* 4 (1986), 214–224.

[15] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 11, Feb (2010), 625–660.

[16] Martin Fischbach, Dennis Wiebusch, and Marc Erich Latoschik. 2017. Semantic Entity-Component State Management Techniques to Enhance Software Quality for Multimodal VR-Systems. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 4 (2017), 1342–1351. DOI: 10.1109/TVCG.2017.2657098.

[17] Sascha Gebhardt, Sebastian Pick, Franziska Leithold, Bernd Hentschel, and Torsten Kuhlen. 2013. Extended pie menus for immersive virtual environments. *IEEE transactions on visualization and computer graphics* 19, 4 (2013), 644–651.

[18] Jawaid A. Ghani and Satish P. Deshpande. 1994. Task characteristics and the experience of optimal flow in human—computer interaction. *The Journal of Psychology* 128, 4 (1994), 381–391.

[19] Enscape GmbH. 2019. *Enscape*. Retrieved April 03, 2019 from https://enscape3d.com/

[20] Carrie Heeter. 1992. Being there: The subjective experience of presence. *Presence: Teleoperators & Virtual Environments* 1, 2 (1992), 262–271.

[21] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[22] Jörn Hurtienne and Anja Naumann. 2010. QUESI—A questionnaire for measuring the subjective consequences of intuitive use. *Interdisciplinary College* 536 (2010).

[23] IrisVR Inc. 2019. *IrisVR*. Retrieved April 03, 2019 from https://irisvr.com/

[24] IrisVR Inc. 2019. *Why This Public Utility Company Went from Unity & Unreal to Prospect for VR (And How Much It Saved Them)*. Retrieved April 27, 2019 from https://blog.irisvr.com/navisworks-vr-case-study

[25] ISO Central Secretary. 2018. *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. Standard ISO 9241-11:2018. International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/62711.html

[26] Seung-A A. Jin. 2011. "I feel present. Therefore, I experience flow:" A structural equation modeling approach to flow and presence in video games. *Journal of Broadcasting & Electronic Media* 55, 1 (2011), 114–136.

[27] Seung-A A. Jin. 2012. "Toward integrative models of flow": Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *Journal of Broadcasting & Electronic Media* 56, 2 (2012), 169–186.

[28] Michael Johnston, Philip R Cohen, David McGee, Sharon L Oviatt, James A Pittman, and Ira Smith. 1997. Unification-based multimodal integration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 281–288.

[29] Yea-Seul Kim, Mira Dontcheva, Eytan Adar, and Jessica Hullman. 2019. Vocal Shortcuts for Creative Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 332, 14 pages. https://doi.org/10.1145/3290605.3300562

[30] Gregory Kramida. 2016. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE transactions on visualization and computer graphics* 22, 7 (2016), 1912–1931.

[31] Marc Erich Latoschik. 2001. A General Framework for Multimodal Interaction in Virtual Reality Systems: PrOSA. In *The Future of VR and AR Interfaces - Multimodal, Humanoid, Adaptive and Intelligent. Proceedings of the Workshop at IEEE Virtual Reality 2001 (GMD report)*, Wolfgang Broll and Leonie Schäfer (Eds.). 21–25. http://trinity.inf.uni-bayreuth.de/download/A_General_Framework_for_Multimodal.pdf

[32] Marc Erich Latoschik. 2002. Designing Transition Networks for Multimodal VR-Interactions Using a Markup Language. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. IEEE Computer Society, 847719, 411.

[33] Marc Erich Latoschik. 2005. A user interface framework for multimodal VR interactions. In *Proceedings of the 7th international conference on Multimodal interfaces*. ACM, 1088479, 76–83.

[34] Marc Erich Latoschik, Martin Fröhlich, Bernhard Jung, and Ipke Wachsmuth. 1998. Utilize Speech and Gestures to Realize Natural Interaction in a Virtual Environment. In *IECON'98: Proceedings of the 24th*

annual Conference of the IEEE Industrial Electronics Society, Vol. 4. 2028–2033. http://trinity.inf.uni-bayreuth.de/download/usg_to_realize.pdf

[35] Joseph J. LaViola Jr, Ernst Kruijff, Ryan P. McMahan, Doug Bowman, and Ivan P. Poupyrev. 2017. *3D user interfaces: theory and practice*. Addison-Wesley Professional.

[36] Anja Naumann and Jörn Hurtienne. 2010. Benchmarks for intuitive interaction with mobile devices. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. ACM, 401–402.

[37] Sharon Oviatt. 2003. Advances in robust multimodal interface design. *IEEE computer graphics and applications* 5 (2003), 62–68.

[38] Sharon Oviatt and Philip Cohen. 2000. Perceptual User Interfaces: Multimodal Interfaces That Process What Comes Naturally. *Commun. ACM* 43, 3 (March 2000), 45–53. https://doi.org/10.1145/330534.330538

[39] Daniel Paes, Eduardo Arantes, and Javier Irizarry. 2017. Immersive environment for improving the understanding of architectural 3D models: Comparing user spatial perception between immersive and traditional virtual reality systems. *automation in Construction* 84 (2017), 292–303.

[40] Eeva M. Pilke. 2004. Flow experiences in information technology use. *International journal of human-computer studies* 61, 3 (2004), 347–357.

[41] Michael I. Posner. 1980. Orienting of attention. *Quarterly journal of experimental psychology* 32, 1 (1980), 3–25.

[42] Ronald Rosenfeld, Dan Olsen, and Alex Rudnicky. 2001. Universal speech interfaces. *interactions* 8, 6 (2001), 34–44.

[43] Marc A. Schnabel and Thomas Kvan. 2003. Spatial understanding in immersive virtual environments. *International Journal of Architectural Computing* 1, 4 (2003), 435–448.

[44] Ben Shneiderman. 2007. Creativity support tools: Accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

[45] Ben Shneiderman, Gerhard Fischer, Mary Czerwinski, Mitch Resnick, Brad Myers, Linda Candy, Ernest Edmonds, Mike Eisenberg, Elisa Giaccardi, and Tom Hewett. 2006. Creativity support tools: Report from a US National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction* 20, 2 (2006), 61–77.

[46] Ivan E. Sutherland. 1968. A head-mounted three-dimensional display. In *Proceeding of the Fall Joint Computer Conference. AFIPS Conference Proceedings. (AFIPS)*, Vol. 33. Arlington, VA, 757–764.

[47] Russell M. Taylor II, Thomas C Hudson, Adam Seeger, Hans Weber, Jeffrey Juliano, and Aron T. Helser. 2001. VRPN: a device-independent, network-transparent VR peripheral system. In *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 55–61.

[48] Unity Technologies. 2017. *Unity*. Retrieved October 25, 2018 from https://unity3d.com/

[49] Robert Tscharn, Marc Erich Latoschik, Diana Löffler, and Jörn Hurtienne. 2017. "Stop over There" – Natural Gesture and Speech Interaction for Non-Critical Spontaneous Intervention in Autonomous Driving. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI)*. 91–100.

[50] Karl T. Ulrich and Steven D. Eppinger. 2004. Product architecture. *Product design and development* 3 (2004), 163–186.

[51] VRTK. 2018. *VRTK - Virtual Reality Toolkit*. Retrieved October 25, 2018 from https://www.vrtk.io/

[52] Dennis Wiebusch, Chris Zimmerer, and Marc Erich Latoschik. 2017. Cherry-Picking RIS Functionality – Integration of Game and VR Engine Sub-Systems based on Entities and Events. In *10th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS)*. IEEE Computer Society.

[53] Maliha Zaman, Murugan Anandarajan, and Qizhi Dai. 2010. Experiencing flow with instant messaging and its facilitating role on creative behaviors. *Computers in Human Behavior* 26, 5 (2010), 1009–1018.

[54] Ferdinand Rudolf Hendrikus Zijlstra. 1993. Efficiency in work behaviour: A design approach for modern tools. (1993).

[55] Chris Zimmerer, Martin Fischbach, and Marc Latoschik. 2018. Semantic Fusion for Natural Multimodal Interfaces using Concurrent Augmented Transition Networks. *Multimodal Technologies and Interaction* 2, 4 (2018), 81.

[56] Chris Zimmerer, Martin Fischbach, and Marc Erich Latoschik. 2018. Concurrent Augmented Transition Network — Project Page. https://www.hci.uni-wuerzburg.de/projects/mmi/. Last accessed 2018-08-22.