Inclusive Voice Interaction Techniques for Creative Object Positioning

Farkhandah Aziz
DMT Lab, Birmingham City University, Birmingham, United Kingdom, farkhandah.komal@mail.bcu.ac.uk

Chris Creed
DMT Lab, Birmingham City University, Birmingham, United Kingdom, chris.creed@bcu.ac.uk

Maite Frutos-Pascual
DMT Lab, Birmingham City University, Birmingham, United Kingdom, maite.frutos@bcu.ac.uk

Ian Williams
DMT Lab, Birmingham City University, Birmingham, United Kingdom, ian.williams@bcu.ac.uk

ABSTRACT

People with physical impairments can experience challenges when using traditional input devices such as a mouse and keyboard for creative visual design work. Speech interaction is an alternative input method that can facilitate a more inclusive approach for supporting creative work. However, there has been a lack of work to date investigating how to position and move creative assets (e.g. images) around a digital workspace via speech interaction techniques. We present three multimodal speech interaction approaches to support the positioning of graphical objects around a design canvas via speech commands: Speed-based Control, Location Guides, and Positional Guides. A user evaluation with non-disabled participants (*N*=30) found that the Location Guides approach was significantly more efficient, accurate, and usable for the positioning of images when compared with the other methods. A follow-up study with physically impaired users (*N*=6) demonstrated they were able to effectively position images around a design canvas using the Location Guides technique with participants also rating this approach as exhibiting a high level of usability.

CCS CONCEPTS • Human-centered computing → Accessibility; Accessibility design and evaluation methods.

KEYWORDS: Assistive Technology; Multimodal Input; Voice Interaction; Object Manipulation; Inclusive Design.

1 INTRODUCTION

People with physical impairments can experience significant barriers when attempting to produce creative work using mainstream visual design applications [7, 8] such as Adobe Photoshop [1], Illustrator [2], and XD [3]. These challenges typically arise due to the use of traditional input tools (i.e. a mouse and keyboard) being the predominant interaction paradigm within creative software [6, 12, 17, 35]. The use of speech input to control systems holds significant potential to make visual design work more accessible, although there has been little research in this area to date. Previous studies have explored the potential of speech interaction to support people with physical impairments in creative drawing activities [28, 29]. However, these studies represent initial

investigations and there remain several key areas that require further research to understand more deeply the potential of this technology in a creative domain. One crucial area where there has been a lack of work to date is around how graphical objects can efficiently be positioned around a digital canvas. A traditional mouse and keyboard input combination enables users to rapidly move digital objects around a design canvas such as nudging an object into a specific position or snapping to an alignment guide [9, 21]. This is an essential component of controlling graphical design interfaces, yet it is unclear how this could be facilitated via voice interaction.

We address the lack of research in this area through developing and investigating new interaction techniques to support people with physical impairments in positioning graphical objects around a digital canvas. By "physical impairments", we primarily refer to people who experience barriers in using traditional input devices (e.g. people with motor neurone disease, cerebral palsy, repetitive strain injury, tremors, etc. where speech is not also significantly impaired), as well as certain forms of situational and temporary physical impairments. We present three different multimodal speech approaches – Speed-based Control (using transformation speed and simple voice commands such as "left", "right", etc.), Location Guides (involving the use of positional labels), and Positional Guides (where traditional guidelines support object placement). Each approach also utilises switch input (e.g. a mechanical switch, head tracker, foot pedal, keyboard and other assistive tools) for initiating the speech recogniser. We present results from an initial evaluation with non-disabled participants (N=30) demonstrating that the Location Guides approach was more efficient, accurate, and usable in positioning objects than Speed-based Control and Positional Guides. After iterative updates the Location Guides method was evaluated in a follow-up study with physically impaired participants (N=6). Results from this study validated findings from the first evaluation and demonstrated how users with physical impairments can effectively position objects around a digital canvas via a multimodal speech input.

This work therefore presents three primary contributions: (1) three multimodal speech interaction approaches for positioning graphical objects, (2) a user evaluation presenting new insights around the use of multimodal speech interaction for object manipulation, and (3) validation of the Location Guides approach in supporting people with physical impairments to efficiently control objects on a design canvas.

2 RELATED WORK

Previous research has investigated the potential of speech interaction to support people with physical impairments in producing freeform drawings – for instance, Harada et al. [13, 14, 15] explored the use of a vocal joystick that enables continuous voice input in the form of vowel sounds to guide drawing directions whilst using a digital brush. Van der Kamp and Sundstedt [22] examined the use of voice commands for manipulating drawing tools combined with eye gaze interaction for controlling the mouse cursor and found this approach supported a more efficient drawing process. Moreover, Adobe [40] recently introduced grid numbers and labels to support users in accessing application features via voice commands (e.g. to select drawing tools, properties, layers, and menu items), although this has not been formally evaluated to date within the academic field.

Other studies have also explored multimodal approaches for creating artistic work, image editing, and graphical manipulation – for instance, Laput et al. [26] presented the PIXELTONE application where direct manipulation (via touch) is used to select parts of the image, along with a limited set of high level voice commands to perform image editing operations (e.g. applying filters). Research findings highlighted a preference for the combination of a speech and touch input approach as compared to touch only input for image

processing operations. Srinivasan et al. [32] also presented a similar approach using natural language speech commands and touch input for image editing operations. Touch input was used to select interface elements and voice commands (e.g. “change fill color”, “add a sepia filter”) were suggested to users based on their context. Results highlighted positive perceptions from participants although there were issues with speech recognition during the study. Sedivy and Johnson [30] presented a multimodal approach where speech input is used for performing sketching operations (e.g. drawing shapes, lines, and coloring) whilst a stylus pen was utilized for object selection, rotation, resizing, dragging shapes, and layer navigation. Results from a user study found that speech input saved time in accessing drawing features, although this was an informal evaluation with only small group of non-disabled users. Hauptmann [16] conducted a study to evaluate three different interaction approaches for manipulating graphical objects – these included speech only, gesture only, and a combination of both speech and gesture interaction. Participants were given a series of tasks to move, scale, and rotate a single cubed shaped object presented on the screen display. Results found that the majority of participants preferred the combination of both speech and gesture input for object manipulation. Hiyoshi and Shimazu [17] also presented a multimodal approach where mouse pointing was used to specify the target position and voice was used for drawing and manipulating basic shapes (e.g. via statements such as “place the object here”). Moreover, Kim et al. [25] investigated the use of short vocal commands in creative applications to support expert designers and found they can help creative experts to access various design features more efficiently, thus reducing cognitive and physical load.

The work highlighted demonstrates the wider potential of speech interaction to facilitate the production of visual design work, although further research is also required to understand the optimal approaches for supporting crucial and common low-level tasks that support an efficient creative workflow. For instance, the ability to effectively position objects around a digital canvas is essential, yet it remains unclear how this should be achieved via voice-based interaction. Moreover, the production of creative visual design work via speech interaction holds much potential for people with physical impairments, although there have also been a lack of studies evaluating this approach with disabled users. Further research is therefore required to thoroughly explore different interaction techniques that facilitate fundamental visual design activities (such as object positioning) to ensure optimal approaches are developed that support the creative process for people with physical impairments.

3 PROTOTYPE DESIGN

We developed a new research prototype to investigate different object positioning approaches within a design canvas via multimodal interaction. The application was built using HTML, CSS, and JavaScript – with the Web Speech API [41] used for detecting speech input from users. The prototype simulates a visual design application and displays a portfolio website design for a fictional professional photographer (Figure 1).

The main design canvas (Fig. 1 (a)) is fixed to 700x560 pixels to ensure that it can run on modern browsers without the need for vertical or horizontal scrolling (to help avoid inconsistent user experiences during evaluation studies). The design canvas consists of nine thumbnail sized images (150x90 pixels) which is similar in size and dimensions to those used in mainstream social media platforms (e.g. Facebook [42], LinkedIn [43], and Instagram [44]) to display posts and profiles pictures. One of these images is the interaction object (Fig. 1 (b)) that is presented in the top right corner of the design canvas (at the start of a task) and can be moved in all directions using a range of speech commands. The target placeholder (Fig. 1 (c)) is displayed as a grey box

which represents the final position where an interaction object needs to be positioned. The placeholder is the same size as the interaction object (150x90 pixels) to ensure the image can be accurately placed over it. The speech command panel (Fig. 1 (d)) is displayed at the top middle section of the screen to help users visualize spoken voice commands which the system has recognized. Switch input (e.g. a keyboard, mechanical switch, head tracker, foot pedal, etc.) is utilised for initiating the speech recogniser on each task screen across three interaction approaches. Audio feedback (a popping sound effect) is also played after a voice command has been issued to make the user aware that their input has been recognized.

The transformation speed (Fig. 1 (e)) can be seen in the top right corner and is used to facilitate users in moving interaction objects slower or faster depending on a user's preferences. The transformation speed can range from values of "1" to "500" and is controlled using the voice command 'speed' followed by a number (i.e. "speed 10"). For instance, when a user wants to move the interaction object towards left and the transformation speed is set to 10, the image will move 10 pixels to the left when the appropriate voice command is issued (i.e. "left"). The transformation speed can also be used to fine-tune the interaction object position when it is closer to the target placeholder (i.e. selecting the minimum speed value, e.g. "speed 1"). The interaction object is moved based on pixel values as opposed to continuous animation at the set speed. This decision was taken as latency in processing of speech recognition can result in slight delays of commands being issued, which in turn can lead to objects moving beyond the user's intended target position (if continuous animation is used) [23].

The supported speech commands (Fig. 1 (f)) at the bottom of the canvas are always visible to help users in recalling the available commands – these commands are dynamically changed on different screens depending on the object positioning approach used. The following subsections present the three different object positioning approaches – Speed-based Control, Speed Control + Location Guides, and Speed Control + Positional Guides.

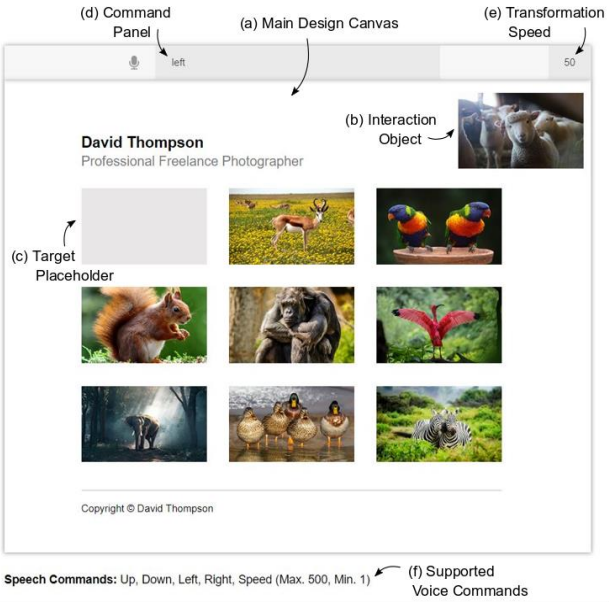

Figure 1: The research prototype interface showing (a) main design canvas, (b) interaction object, (c) target placeholder, (d) speech commands panel, (e) transformation speed, and (f) supported voice commands.

3.1 Speed-based Control

This approach uses simple voice commands to move objects (images) around the design canvas. These commands include “left”, “right”, “down”, and “up” to move the interaction object presented in the top right position of the design canvas (Figure 1). These voice commands are informed and motivated from previous work in the field (e.g. [24]) where directional voice commands were used to control different interface elements (e.g. a mouse cursor). This approach mainly relies on the transformation speed (as discussed above) to move and position the interaction object over the relevant placeholder. There are no additional visual cues or layout tools (in contrast to other two approaches) to support the image movement around the design canvas.

3.2 Speed Control + Location Guides

This approach utilizes the same speech commands used in the “Speed-based Control” approach, along with the use of location guidance to assist image positioning around the design canvas. The location guides were presented as a grid of circular labels (numbered from 1-90) overlaid on top of design canvas. The size of the circular labels was set to 20x20 pixels and were placed 75px (≈2cm) apart from each other. A series of informal usability tests evaluating a range of different label sizes and distances helped to inform a balance between sizes, numbers, and distancing of location guides to ensure users are not overloaded with options (which could result in a cluttered user experience). The opacity of the interface elements is lowered when the location guides are displayed to ensure the grid of labels are not obscured by visual content on the canvas. This approach is similar to the method used in Adobe XD [40] where labels were used for selecting tools and features via voice, although this type of approach has not yet been explored for object positioning. The location guides can be displayed through saying “locations” – objects can then be moved saying the number contained within the label that is closest to target placeholder thus resulting in the top-left corner of the interaction object being placed over the appropriate location guide (Figure 2). Location guides can be hidden stating “hide” command – users can then use the Speed-based Control method and transformation speed to refine specific location of the object.

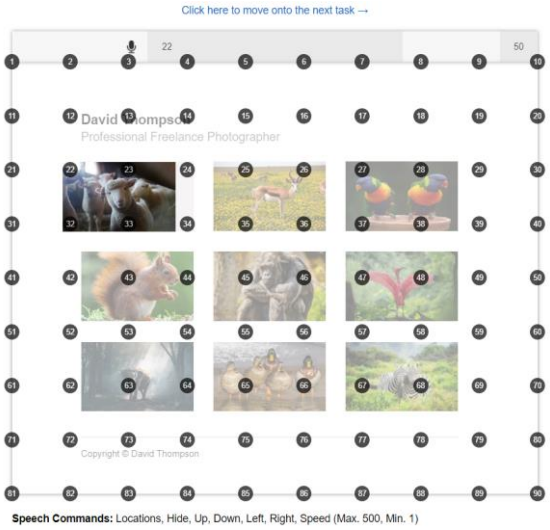

Figure 2: An example of the Location Guides feature where the top left corner of an image has been placed at the closest possible position (Location 22) over the target placeholder.

3.3 Speed Control + Positional Guides

The third approach uses standard vertical and horizontal positional guidelines commonly used in mainstream design software (e.g. Adobe Photoshop [1] and Illustrator [2]). A similar approach has also been used in related work investigating the positioning of graphical objects within a design canvas via the combination of gaze interaction and mechanical switches to support people with physical impairment [9]. We adapted these approaches for use via voice control (Figure 3) – users can initially enable the positional guide feature by stating "guides". A horizontal and vertical line are then displayed within the main canvas which can be moved through stating "left", "right", "up" or "down". The transformation speed is also associated with moving guidelines and can be changed using the same speed commands as discussed previously. As a first step, the vertical and horizontal lines need to be moved so that their intersection point is at the top left corner of the user's desired location. Use of the "snap" command then moves the interaction object to the point at which the lines intersect (Figure 3). The "hide" command can be used to hide the positional guides from the main view. Users can also utilize the Speed-based Control if the interaction object is not snapped at the exact position over the relevant placeholder.

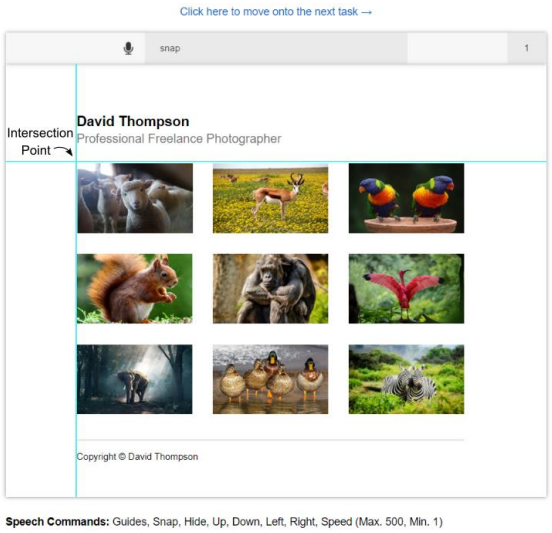

Figure 3: An example of the Positional Guides feature with an image (top left) snapped at the intersection of the lines using the "snap" command.

4 USER EVALUATION

An online user evaluation was conducted with non-disabled participants to investigate whether they were able to position objects effectively via multimodal interaction. A within participant design was utilised with non-disabled participants to ensure the approaches were viable and usable prior to running evaluations with physically impaired participants. It was felt that this first study would provide interesting insights into the different approaches and opportunities for iterative improvements prior to working with disabled participants.

4.1 Participants

Thirty non-disabled participants (15 male and 15 female) were recruited from a population of University staff and students. Participants were aged between 19 to 52 years (Mean = 28.13, SD = 7.57) and were native-English speakers. They were assessed based on their level of experience with graphical design software (10 Novice, 17 Intermediate, and 3 Experts), prototyping applications (15 Novice, 13 Intermediate, 2 Experts), and speech interaction technology (8 Novice, 16 Intermediate, and 6 Experts).

4.2 Apparatus and Procedure

Participants were required to use their own computer and microphone for voice input, as well as a keyboard for simulating a switch to control the speech recognizer (using spacebar key). The Google Chrome browser was required for experimental tasks due to browser compatibility with the Web Speech API [41]. Participants initially communicated via email to confirm the testing schedule as well as the preferred platform for the evaluation session (e.g. Microsoft Teams [27] or Zoom [38]). At the start of testing sessions the URL of the object positioning prototype was provided to participants which they were asked to access and then share their screen content. A researcher then provided pre-test instructions to ensure participants understood the purpose of the study – once this was confirmed, they redirected to a consent page, followed by a pre-test survey asking questions around demographic information, graphic design and speech interaction experience.

Participants then completed training tasks (i.e. moving a single object around a blank design canvas) to understand how to operate the relevant object positioning method before starting the main tasks for each interaction approach. Participants then clicked a link at the top of the screen to move onto the main evaluation which consisted of three interaction approaches tested using nine object positioning tasks (i.e. 27 tasks in total). The order of interaction modes and tasks were randomized to minimize the potential for order bias. For each task, a different image was always placed in the top-right corner of the interface to be positioned over its associated placeholder (i.e. the grey box). Placing all interaction images in the top-right helped to ensure a consistent approach was adopted for comparing the different interaction techniques, as well as ensuring that participants had to move images in all directions to successfully complete tasks. Before starting a task participants activated the speech recogniser on each task screen using spacebar key. After successfully moving and positioning an interaction object over the relevant placeholder participants selected the next task link above the design canvas.

Once all nine tasks had been completed for an interaction technique participants were presented with the SUS form [5] to complete. They then started on the next technique with an initial training session, followed by the main tasks, and then the SUS form again. Participants were also asked eight open-ended questions at the end of testing session to explore their perceptions of the different interaction techniques. Testing sessions lasted between 26 to 60 minutes in total.

4.3 Measures

Task completion time, positional accuracy, speech recognition performance, and SUS scores [5] were calculated to evaluate the three interaction techniques. Task completion time was measured from when participants selected the start task link until they selected the next task link upon task completion. Positional accuracies (distances) were measured through task-wise arrangement of final interaction objects and target

placeholders locations, calculating the differences between these values, and then finding the Euclidean distance values [33] from *x* and *y* values for each interaction approach.

Speech recognition performance was measured via speech recognition errors where "errors" were identified as the recognizer incorrectly interpreting a voice command (e.g. the speech command "right" misrecognized as "write"). SUS [5] was used to evaluate perceptions of usability for each interaction approach. Post-study questions also explored participants' perceptions around each interaction approach, their overall impressions of using multimodal speech interaction for object positioning, and suggestions for improvements.

4.4 Results

The Shapiro-Wilk's [31] normality test ($p > 0.05$) found task completion data to be normally distributed while positional accuracy, speech performance, and SUS data were not normally distributed. A one-way repeated measure ANOVA [11] was utilized to analyze the differences between task completion times for each interaction approach. We used non-parametric Friedman test of differences for repeated measures with Bonferroni correction to analyze positional accuracy, speech performance, and SUS scores. Wilcoxon signed rank [37] was used for Post-hoc tests to further analyze the differences in positional accuracy, speech performance, and usability scores.

4.4.1 Task Completion Time.

A statistically significant difference was observed between Speed-based Control (Mean = 11.33, SD = 2.28), Location Guides (Mean = 8.22, SD = 1.92), and Positional Guides (Mean = 12.39, SD = 2.97) in relation to task completion time (F (2, 58) = 49.16, $p < 0.001$, partial $\eta 2 = 0.629$). Post-hoc LSD tests [34] showed a significant difference between Location Guides and Speed-based Control (sig = 0.001, $p < 0.05$), Location Guides and Positional Guides (sig = 0.001, $p<0.05$), and Speed-based Control and Positional Guides (sig = 0.012, $p< 0.05$). Figure 4 demonstrates participants took significantly less time to complete all tasks using the Location Guides approach in comparison to the other two techniques.

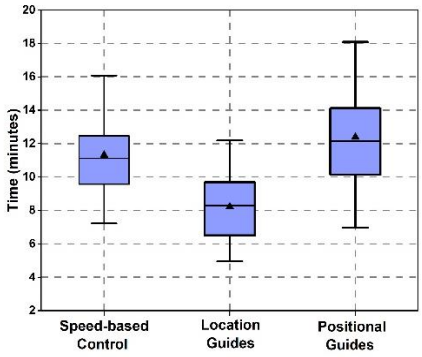

Figure 4: Task completion times

4.4.2 Positional Accuracy.

Average positional accuracy based on average Euclidean distance values across the Speed-based Control method was 0.88 (SD = 0.81), Location Guides 0.62 (SD = 0.49), and Positional Guides was 0.68 (SD = 0.89).

Friedman test results highlighted significant differences in positional accuracy (X^2 = 0.014, *df* = 2, $p < 0.05$). The post-hoc Wilcoxon signed rank showed a significant difference in positional accuracy between Location Guides and Speed-based Control (Z= -3.92, $p < 0.001$) and Positional Guides and Speed-based Control (Z= -3.58, $p < 0.001$). No significant differences were found between Location Guides and Positional Guides (Z = -0.69, $p = 0.48$). Figure 5 shows the positional accuracy as average positional distances which demonstrates how close (accurate) the interaction objects were placed over their target placeholders.

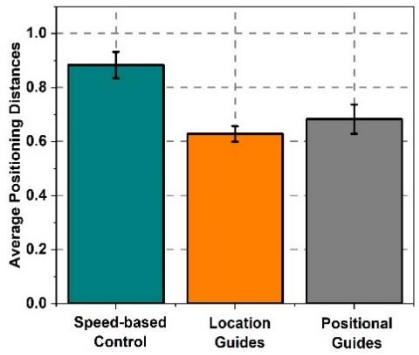

Figure 5: Positional accuracy

4.4.3 Speech Performance.

The total vocal commands issued across all 30 participants for Speed-based Control was 4596 (SD = 0.81), 3941 (SD = 0.49) for Location Guides, and 4837 (SD = 0.89) for Positional Guides. The total speech recognition errors for Speech-based Control was 376 (8.2%), 321 (8.14%) for Location Guides, and 437 (9.03%) for Positional Guides. Friedman test and post-hoc Wilcoxon signed rank showed no significant differences in speech performance among three speech interaction approaches (Z = -1.72, $p = 0.09$; Z = -1.17, $p = 0.24$; Z = -1.87, $p = 0.06$).

4.4.4 Usability Evaluation.

The average SUS scores for Location Guides was 86.56 (SD = 14.09) and is rated as "excellent" according to Bangor et al. [5]. Speed-based Control (Mean = 76.83, SD = 14.99) and Positional Guides (Mean = 80.25, SD =14.27) can also be labelled as exhibiting "good" usability [5]. Significant differences were found across the three interaction approaches (X^2 = 0.001, *df* = 2, $p < 0.05$). The post-hoc Wilcoxon signed rank found a significant difference between Location Guides and Speed-based Control (Z = -3.03, $p < 0.001$). A significant difference was also observed between Location Guides and Positional Guides (Z = -2.42, $p < 0.001$). No significant differences were found between Positional Guides and Speed-based Control (Z = -1.33, p = 0.19).

4.4.5 Qualitative Evaluation.

Overall, twenty participants preferred Location Guides, seven preferred Positional Guides, while three participants preferred Speed-based Control. Participants emphasized that the Location Guides approach was the most efficient and easy to use over the other methods *"…location guides is easy to understand, and quickest, as you are able to get close to the position you want then just fine-tune at that location"* (Participant

13). Overall, all participants provided positive feedback in terms of using the three multimodal object positioning techniques. Participant 24 commented that they *"… liked all methods, straightforward to use, not much training required, pretty intuitive"*. Four participants commented that the system occasionally interpreted their utterances incorrectly – for instance, the "right" command was identified as "write" or "wright", "left" as "lift", and "snap" as "nap". Participant 6 commented that the speech command panel (i.e. (Fig. 1 (f)) currently displays everything that a user verbalizes (including general conversational speech unrelated to controlling the application) and suggested filtering this to only display recognised commands. No participants highlighted speech recognition or the use of a keyboard to control the speech recognizer as a significant issue in completing the evaluation tasks.

4.5 Study Discussion

The Location Guides approach was found to be significantly faster and more usable for the image positioning tasks than Positional Guides and Speed-based Control. Participants were also more accurate in positioning images when using Location Guides as compared to Speed-based Control (although there were no significant differences between Location and Positional Guides). Subjective feedback also highlighted that the majority of participants preferred the Location Guides method and provided positive feedback in terms of its efficiency and usability in placing images around the canvas. It was therefore decided that the Location Guides approach would be taken forward for further development and evaluation work in partnership with disabled participants.

5 FOLLOW-UP STUDY

Iterative updates were made to the research prototype based on the feedback received from the first study – in particular, the speech recognition system performance was improved through mapping the homophones detected in the first study with the relevant "correct" commands (e.g. "write" was added to work similar to "right", etc.). We also applied filtering on recognized speech input to ensure that only supported voice commands were displayed in the command panel. Since the Location Guides method had a high SUS score and was considered easy to use and efficient, no further major updates were made to the prototype.

5.1 Participants

Six participants (3 male and 3 female) were recruited through online advertisements and the London RSI support group [39]. Participants were aged between 27 to 49 years (Mean = 38.33, SD = 7.69) and all were native-English speakers. Table 1 details participants' impairments, technical experience, multimodal tools used during the testing session, and SUS scores.

5.2 Apparatus and Procedure

Institutional Review Board approval was obtained for the research study. All participants used their own microphones for voice control and were encouraged to use their chosen form of switch input (e.g. foot pedal, mechanical switch, head tracker, keyboard, etc.). Evaluations were conducted online using the Zoom video conferencing platform [38] and participants had to complete the same nine image positioning tasks using only the Location Guides method. The same procedure was followed for this study as discussed in Section 4.2. A semi structured interview was also conducted to investigate participants' experience and perceptions of the positioning technique. Evaluation sessions lasted between 20-35 minutes.

5.3 Results

5.3.1 Usability, Task Completion Time, Positional Accuracy, Speech Performance.

An average SUS score of 85.42 (SD = 12.28) was received for the Location Guides method which can be labelled as "Excellent" [5] (Table 1). The overall task completion times ranged between 6.24 (P2) – 13.95 minutes (P3), whilst the average task completion time across all six participants was 10.93 (SD = 3.18). The positional accuracy was calculated using Euclidean distances [33] for individual participants where average values ranged from 0.54 – 1.22 (Mean = 0.84, SD = 0.25). A total of 597 voice commands were issued across all six participants where 41 (6.86%) speech recognition errors were identified.

Table 1: Participants Details: RSI = Repetitive Strain Injury; MM = Muscular Myopathy; GD = Graphical Design (Software); IP = Interface Prototyping (Software); ST = Speech Technology; AT = Assistive Technology

ID	Age/ Gender	Physical Impairments	Condition Details	Technical Experience	Multimodal tool used during testing session	SUS Score
P1	27 (M)	Tenosynovitis (Since July 2020)	Wrist Pain; Joint swelling and stiffness; Difficulty in using fingers.	*GD*: Average; *IP*: Novice; *ST*: Dragon, Apple Siri; *AT*: Head Tracker, Foot pedal.	Speech + Foot pedal	92.50
P2	34 (M)	RSI (Since 2014)	Hand tremors; Shooting pain in hands and arms; Pain in wrists; Tingling sensation in fingers.	*GD*: Novice; *IP*: Novice; *ST*: Dragon software, Talon, Apple Siri, and Google Assistant *AT*: Eye tracking, Vertical mouse.	Speech + Keyboard	97.50
P3	49 (F)	RSI (Since 2012)	Severe pain and discomfort in hands; Tiredness in shoulders and upper arms.	*GD*: Novice ; *IP*: Novice; *ST*: Dragon software; *AT*: Novice.	Speech + Keyboard	82.50
P4	41 (F)	RSI (Since 2010)	Fatigue; Shoulder pain; Sore wrists occasionally; Pulsing pain in fingers.	*GD*: Expert ; *IP*: Expert; *ST*: Dragon software; *AT*: Vertical mouse, Mechanical Switch.	Speech + Jelly Bean Switch	100
P5	33 (M)	MM (Since 2009)	Muscles weakness; fatigue; Lack of balance; Difficulty with walking without sticks;	*GD*: Average; *IP*: Average; *ST*: Apple Siri; Google Assistant *AT*: Novice.	Speech + Keyboard	67.50
P6	46 (F)	RSI (Since 2000)	Wrist pain, Pain in shoulders and upper arms; tiredness.	*GD*: Novice; *IP*: Novice; *ST*: Dragon software *AT*: Novice.	Speech + Keyboard	72.50

5.3.2 Qualitative Feedback.

All participants provided positive feedback and were able to utilize the features within the research prototype for image positioning. Four participants (P1, P2, P3, and P6) commented that they liked the grid presentation of the location guides and that they enabled them to efficiently position objects ("…*I like that it gives you the control to easily do large imprecise movements followed by doing small more precise adjustments"* [P1]"). Moreover, P3 stated that the Location Guides approach *"… is straightforward and very helpful to align objects because when you are not using mouse moving things around canvas is a difficulty"*. P4 commented that "… *it*

could be an interesting and useful activity for dragging things on the interface for manipulation work". Two participants were also able to effectively utilize their own assistive technologies to complete the tasks (i.e. P1 used a foot pedal for controlling the speech recognizer, whilst P4 used a Jelly Bean switch).

6 DISCUSSION AND FUTURE WORK

This paper has presented a new system that facilitates three different methods of positioning graphical objects around a design canvas via multimodal speech interaction (Speed-based Control, Location Guides, and Positional Guides). The majority of participants from the first study found all three approaches to be viable, although the Location Guides approach was perceived to be more efficient, accurate, and usable. People with physical impairments also provided positive feedback with results demonstrating that they were able to successfully complete image positioning tasks during the follow-up study. This work therefore contributes a deeper understanding around the viability of speech interaction to make core tasks associated with creative activities more accessible. Moreover, the results from this research have wider applicability in other related domains (e.g. positioning of objects in commercial office software – Word processors, presentation applications, operating systems, etc.), as well as the potential to support object positioning on different platforms (e.g. tablets and mobile phones).

One potential limitation of the work is related to the issues observed around accurate speech recognition (which is a known issue within the field [4, 28]), although steps were taken to minimise this impact through using short vocal commands [25], as well as use of homophones and filtering. The positive feedback and usability ratings from participant suggests that this helped to create an accessible and usable experience. The study also only involved native English speakers as participants – it will therefore be important in future work to explore further the efficacy of the recognition system using other languages (as well as additional possible factors such as the gender of participants). Another potential limitation is the interaction objects and placeholders used within the study being designed to match the size of images used in common portfolio design tasks (to help create a more realistic design scenario). However, it will be crucial in future studies to investigate a range of different graphical objects (e.g. shapes, typography, smaller/larger objects, longer/shorter distances [9, 18, 20, 36]) and how users find the experience of positioning objects in these different design scenarios. We feel given the positive findings from the research highlighted that the approaches would still provide an accessible experience in these different contexts, although this will be important to validate.

In terms of other future work opportunities, one key underexplored area is the use of natural language vocal commands to support object positioning. It could be that commands such as "down x pixels", "place at the top of object x", or "align horizontally with object x and vertically with object y" would provide users with more flexibility in placing objects around a canvas. There has also been a lack of work around how other common positioning approaches (e.g. simultaneously selecting and snapping multiple objects to alignment guides) can be made accessible via speech control. Furthermore, it is crucial to investigate different voice supported techniques for other common fundamental object transformations such as scaling and rotation to help create inclusive design environments.

Exploring alternative multimodal approaches to support creative work is also a potentially fruitful area – research has examined the use of speech combined with gestures [10,17] and touch input [26, 32], although other methods could also be beneficial (e.g. using gaze for selecting locations within a canvas and speech to confirm movements, thus helping address known issues such as the Midas Touch [19]).

7 CONCLUSION

Our work presents the first empirical evaluations to investigate different multimodal speech interaction techniques (Speed-based Control, Location Guides, and Positional Guides) for supporting people with physical impairments in positioning objects around a design space. An initial evaluation with thirty non-disabled participants found that the Location Guides approach was more efficient, accurate, and usable than the other two approaches presented. A follow-up study validated that the Location Guides approach was also a suitable and effective approach for people with physical impairments when positioning objects within a design canvas. Feedback from participants across both evaluations also highlighted interesting insights for future research that will help to inform the design of more accessible and creative design environments moving forward.

REFERENCES

[1] Adobe. 1999. Photoshop apps - desktop, mobile, and tablet | Photoshop.com. Retrieved May 31, 2021 from https://www.photoshop.com/en

[2] Adobe Inc. 2012. Adobe Illustrator CS6: Industry-leading vector graphics software. Retrieved May 31, 2021 from https://www.adobe.com/uk/products/illustrator.html

[3] Adobe Inc. 2021. Adobe XD | Fast & Powerful UI/UX Design & Collaboration Tool. Retrieved May 31, 2021 from https://www.adobe.com/uk/products/xd.html

[4] Mohammad M Alsuraihi and Dimitris I Rigas. 2007. How effective is it to design by voice? In *Proceedings of HCI 2007: The 21st British HCI Group Annual Conference*.1–4. DOI:https://doi.org/10.14236/ewic/hci2007.42

[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of usability studies.* 114–123.

[6] Philip R Cohen and Sharon L Oviatt. 1995. The role of voice input for human-machine communication. *Proceedings of National Academy of Sciences*. 9921–9927. DOI:https://doi.org/10.1073/pnas.92.22.9921

[7] Chris Creed. 2018. Assistive technology for disabled visual artists: exploring the impact of digital technologies on artistic practice. *Disability and Society.* 1103–1119. DOI:https://doi.org/10.1080/09687599.2018.1469400

[8] Chris Creed, Russell Beale, and Paula Dower. 2014. Digital tools for physically impaired visual artists. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 253–254. DOI:https://doi.org/10.1145/2661334.2661386

[9] Chris Creed, Ian Williams, and Maite Frutos-Pascual. 2020. Multimodal Gaze Interaction for Creative Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. DOI:https://doi.org/10.1145/3313831.3376196

[10] Carlos Duarte and Joana Neca. 2011. Evaluation of Gestural Interaction with and without Voice Commands. In *Proceedings of IHCI 2011-IADIS conference on Interfaces and Human Computer Interaction*. Retrieved January 27, 2021 from https://www.researchgate.net/publication/256007049

[11] Ellen R. Girden. 1992. ANOVA: Repeated Measures. Google Books. Retrieved May 31, 2021 from https://books.google.co.uk/books?hl=en&lr=&id=JomGKpjnfPcC&oi=fnd&pg=PP7&dq=one+way+repeated+measures+anova&ots=myUtDhTn7B&sig=hIqzBTLl1sivUonLLYkVTnn-Ypg&redir_esc=y#v=onepage&q=one way repeated measures anova&f=false. Sage

[12] Arno Gourdol, Laurence Nigay, Daniel Salber, and Joelle Coutaz. 1992. Two Case Studies of Software Architecture for Multimodal Interactive Systems: VoicePaint and a Voice-enabled Graphical Notebook. *Engineering for Human-Computer Interaction*. 271–84.

[13] Susumu Harada, T Scott Saponas, and James A Landay. 2007. VoicePen: Augmenting pen input with simultaneous non-linguistic vocalization. In *Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI'07*. 178–185. DOI:https://doi.org/10.1145/1322192.1322225

[14] Susumu Harada, Jacob O. Wobbrock, and James A. Landay. 2007. VoiceDraw: A hands-free voice-driven drawing application for people with motor impairments. In *ASSETS'07: Proceedings of the Ninth International ACM SIGACCESS Conference on Computers and Accessibility*. 27–34. DOI:https://doi.org/10.1145/1296843.1296850

[15] Susumu Harada, Jacob O. Wobbrock, Jonathan Malkin, Jeff A. Bilmes, and James A. Landay. 2009. Longitudinal study of people learning to use continuous voice-based cursor control. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* 347–356. DOI:https://doi.org/10.1145/1518701.1518757

[16] Alexander G. Hauptmann. 1989. Speech and gestures for graphic image manipulation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 241–245. DOI:https://doi.org/10.1145/67449.67496

[17] Mayumi Hiyoshi and Hideo Shimazu. 1994. Drawing pictures with natural language and direct manipulation. In *COLING 1994 Volumn 2: The 15th International Conference on Computational Linguistics*. DOI:https://doi.org/10.3115/991250.991262

[18] Ruimin Hu, Shaojian Zhu, Jinjuan Feng, and Andrew Sears. 2011. Use of speech technology in real life environment. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, Berlin, Heidelberg. 62–71. DOI:https://doi.org/10.1007/978-3-642-21657-2_7

[19] Robert J.K. Jacob and Keith S. Karn. 2003. Eye Tracking in human-computer interaction and usability research: Ready to deliver the promises*. In The Mind's eye*. North-Holland. 573–605. DOI:https://doi.org/10.1016/B978-044451020-4/50031-1

[20] Hesham M Kamel and James A Landay. 2000. A study of blind drawing practice: creating graphical information without the visual channel. In *Proceedings of the fourth international ACM conference on Assistive technologies*. 34–41.

[21] Hesham M Kamel and James A Landay. 2002. Sketching images eyes-free: A grid-based dynamic drawing tool for the blind. In *Proceedings of the fifth international ACM conference on Assistive technologies*. 33–40.

[22] Jan Van der Kamp and Veronica Sundstedt. 2011. Gaze and voice controlled drawing. In *Proceedings of the 1st conference on novel gaze-controlled applications.* 1–8. DOI:https://doi.org/10.1145/1983302.1983311

[23] Andrew Sears, Min Lin, and Azfar S. Karimullah. 2002. Speech-based cursor control: understanding the effects of target size, cursor speed, and command selection. *Universal Access in the Information Society.* 30–43. DOI:https://doi.org/10.1007/s10209-002-0034-6

[24] Azfar S Karimullah and Andrew Sears. 2002. Speech-based cursor control. In *Proceedings of the fifth international ACM conference on Assistive technologies*. 178–185. DOI:https://doi.org/10.1145/638249.638282

[25] 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*. 1–14. DOI:https://doi.org/10.1145/3290605.3300562

[26] Gierad Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2185–2194. DOI:https://doi.org/10.1145/2470654.2481301

[27] Microsoft. 2021. Microsoft Teams | Group Chat, Team Chat & Collaboration. *Microsoft Teams*. Retrieved May 31, 2021 from https://www.microsoft.com/en-gb/microsoft-teams/group-chat-software

[28] Takuya Nishimoto, Nobutoshi Shida, Tetsunori Kobayashi, and Katsuhiko Shirai. 1995. Improving human interface in drawing tool using speech, mouse and key-board. In *Proceedings 4th IEEE International Workshop on Robot and Human Communication*. 107–112. DOI:https://doi.org/10.1109/roman.1995.531944

[29] Randy Pausch and James H. Leatherby. 1991. An empirical study: Adding voice input to a graphical editor. In *Journal of the American Voice Input/Output Society*. Retrieved January 27, 2021 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.4829

[30] Jana Sedivy and Hilary Johnson. 1999. Supporting creative work tasks: the potential of multimodal tools to support sketching. In *Proceedings of the 3rd conference on Creativity and Cognition*. 42–49.

[31] Samuel S Shapiro and Martin B Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. 591–611. DOI:https://doi.org/10.2307/2333709

[32] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering natural language commands in multimodal interfaces. *In Proceedings of the 24th International Conference on Intelligent User Interfaces*. 661–672. DOI:https://doi.org/10.1145/3301275.3302292

[33] Liwei Wang, Yan Zhang, and Jufu Feng. 2005. On the Euclidean distance of images. *IEEE transactions on pattern analysis and machine intelligence*. 1334–1339. DOI:https://doi.org/10.1109/TPAMI.2005.165

[34] Lynne J Williams and Herve Abdi. 2010. Fisher's Least Signiflcant Difierence (LSD) Test. *Encyclopedia of research design*. 840–853. Retrieved January 27, 2021 from https://www.researchgate.net/publication/242181775

[35] Shaojian Zhu, Yao Ma, Jinjuan Feng, and Andrew Sears. 2009. Speech-Based Navigation: Improving Grid-Based Solutions. In *IFIP conference on Human-Computer Interaction*. 50–62. Springer, Berlin, Heidelberg

[36] Shaojian Zhu, Andrew Sears, and Jinjuan Feng. 2010. Investigating grid-based navigation:The impact of physical disability. *ACM Transactions on Accessible Computing.* 1–30. DOI:https://doi.org/10.1145/1838562.1838565

[37] Donald W. Zimmerman and Bruno D. Zumbo. 1993. Relative power of the wilcoxon test, the friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, 75–86. DOI:https://doi.org/10.1080/00220973.1993.9943832

[38] Zoom. 2019. Video Conferencing, Web Conferencing, Webinars, Screen Sharing - Zoom. *Zoom* (2019). Retrieved January 25, 2021 from https://www.zoom.us/

[39] Central London RSI Support Group - Home | Facebook. Retrieved January 25, 2021 from https://www.facebook.com/CentralLondonRsiSupportGroup/

[40] Designing Out Loud: Announcing Support for macOS Voice Control in Adobe XD. 2020. Retrieved January 25, 2021 from https://blog.adobe.com/en/2020/02/11/announcing-mac-os-voice-control-adobe-xd.html?scid=fac788f5-fe6f-4be4-a960-871ac58b5f30&mv=social&mv2=owned_social

[41] Using the Web Speech API - Web APIs | MDN. Retrieved May 31, 2021 from https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API/Using_the_Web_Speech_API

[42] Facebook - Log In or Sign Up. Retrieved May 31, 2021 from https://www.facebook.com/

[43] LinkedIn: Log In or Sign Up. Retrieved May 31, 2021 from https://www.linkedin.com/

[44] Instagram. Retrieved May 31, 2021 from https://www.instagram.com/