Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation

ADAM S. WILLIAMS, Colorado State University, USA
FRANCISCO R. ORTEGA, Colorado State University, USA

This research establishes a better understanding of the syntax choices in speech interactions and of how speech, gesture, and multimodal gesture and speech interactions are produced by users in unconstrained object manipulation environments using augmented reality. The work presents a multimodal elicitation study conducted with 24 participants. The canonical referents for translation, rotation, and scale were used along with some abstract referents (create, destroy, and select). In this study time windows for gesture and speech multimodal interactions are developed using the start and stop times of gestures and speech as well as the stoke times for gestures. While gestures commonly precede speech by 81 ms we find that the stroke of the gesture is commonly within 10 ms of the start of speech. Indicating that the information content of a gesture and its co-occurring speech are well aligned to each other. Lastly, the trends across the most common proposals for each modality are examined. Showing that the disagreement between proposals is often caused by a variation of hand posture or syntax. Allowing us to present aliasing recommendations to increase the percentage of users' natural interactions captured by future multimodal interactive systems.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; **Mixed / augmented reality**; **Interaction techniques**; **Empirical studies in HCI**; *User centered design.*

Additional Key Words and Phrases: elicitation, multimodal interaction, augmented reality, gesture and speech interaction

ACM Reference Format:
Adam S. Williams and Francisco R. Ortega. 2020. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 202 (November 2020), 21 pages. https://doi.org/10.1145/3427330

1 INTRODUCTION

Establishing impactful unimodal and multimodal interaction techniques for augmented reality (AR) head-mounted displays (HMDs) starts with understanding unconstrained user behavior. Gesture and speech show promise as the inputs that will be well suited for use in AR-HMDs. Both of these modalities can be tracked with the sensors that come standard on most consumer-available AR-HMDs such as the Microsoft Hololens 2. This minimalism is beneficial. When using AR-HMDs people will likely seek to carry as little extra technology as possible.

Gestures and speech have strengths as both unimodal and multimodal inputs [37]. These strengths have not yet fully been examined. Speech has been found well suited for abstract tasks such as

Authors' addresses: Adam S. Williams, AdamWil@colostate.edu, Colorado State University, 2545 Research Blvd, Fort Collins, Colorado, 80526, USA; Francisco R. Ortega, F.Ortega@colostate.edu, Colorado State University, 2545 Research Blvd, Fort Collins, Colorado, 80526, USA.

2573-0142/2020/11-ART202 $15.00
https://doi.org/10.1145/3427330

multi-object manipulation [49] or selecting a device out of a set of devices [63]. Gestures have been found well suited for direct manipulation [49]. The combination of these modalities can provide a more rich interaction environment than either alone. By understanding the strengths and synergies of these modalities we can better design systems for the end-user.

We can see some of the impacts of new interaction paradigms in the widespread use of multi-touch devices (e.g.,touch screen cell phones) reaching populations that do not commonly use computers but can benefit from the use of technology [18]. Augmented reality is one of the technologies expected to become pervasive in the future, and with that, interactions in AR-HMD environments will become pervasive. Proof AR-HMDs' increased prevalence can been seen in the the United States government's purchase of 100, 000 Microsoft HoloLens 2 units for Army use [9]. There is little standardization for mid-air gestures AR environments [16], the same can be said for speech inputs. Co-occurring gesture and speech interactions, where both gestures and speech are used to convey a message within close temporal proximity of each other, have been analyzed within the context of human to human interaction [26, 33, 36], however, the unconstrained generation of these inputs in human-computer interaction (HCI) has been far less commonly examined [21, 35, 37].

This research presents a study in which participants are tasked with interacting with a virtual object both unimodally and multimodally in an optical see-through AR-HMD environment. These interactions were unconstrained. Gestures, speech, and co-occurring gesture and speech interactions were each tested independently. The main goal of this research was to provide insight on speech interactions, with and without gestures, for object manipulation in AR. To provide robust comparisons, unimodal gesture alone interactions were also examined.

The **contributions** of this research include a detailed analysis of these input modalities' interactions and insights into the changes in those interactions when used multimodally as opposed to unimodally are given. Instead of presenting a single consensus set for each modality, we highlight the common proposals, themes across proposals, and the syntax used for speech interactions. Lastly, timing windows based on the phases of a co-occurring gesture and speech interaction are constructed. Showing that the information content of an interaction is closely aligned with the stroke of a gesture. Based on those findings this paper establishes some guidelines for multimodal gesture and speech input development in this emerging area.

1.1 Motivation

Interactions with systems should be intuitive [41]. One way of achieving that is by leveraging interaction modalities that we are familiar with. Interpersonal communication is rich with gesture and speech interaction [33]. Communication is formed in both gesture and speech channels simultaneously, with each channel impacting the formation of a message by the other channel [26]. Enabling a system to accept gesture and speech as both unimodal and multimodal input channels, is an important step towards creating intuitive augmented reality interaction design.

When participants were given the option to chose modalities, they chose to combine gesture and speech inputs 60% to 70% of the time [15, 19]. This preference can be used to improve recognition [12]. End-users feel that interactions with a system are more natural when they can chose input modalities based on their preference [4, 25]. By leveraging this preference and multimodal inputs, many benefits can be realized. The use of multiple input channels can lead to mutual disambiguation of information lost in the other channel [24, 29, 45], as well as lead to less verbose interactions by allowing for two communication channels to send non-redundant information simultaneously [17]. Gesticulation is closely linked to the structure of co-occurring speech, allowing for better error recovery in recognizers [29].

Optical see-trough AR-HMDs (e.g., Magic Leap One and Microsoft Hololens versions 1 & 2) are starting to implement gesture and speech interactions. That said, these interactions could still use much improvement. Some of the interactions implemented seem built to improve recognition accuracy rather than improving user experience. For example, Magic Leap's C gesture is fairly easy to detect (being a static symbolic gesture) but may not be the most intuitive. Often if gesture sets are not designed with an emphasis on recognition they are designed by experts [62]. User-defined gesture sets have been shown to be up to 24% more memorable [40] and to be preferred to expert-designed gesture sets [61].

This work is not on multimodal fusion (or recognition) [11], rather, it is on multimodal interaction, input generation, and design. Nevertheless, the results of our study can be used by researchers working on multimodal fusion. We use participatory design guidelines to work with potential end-users of AR-HMDs to find what inputs within each modality they would instinctively use [37, 61]. The timing information for phases of a multimodal interaction can help tune recognition windows in multimodal fusion systems. The combination of work on elicitation, such as this study, and multimodal fusion will help HCI build systems with more natural interactions. The technological gap between the feasibility of traditional inputs and gesture with speech inputs is being minimized, soon the later may become more efficient [4]. This work provides information on the top few interaction proposals for each modality, interaction themes across modalities, co-occurring gesture and speech timing information by phase of interaction, and design guidelines on input design for AR building environments.

2 PREVIOUS WORK

2.1 Gesture Elicitation

Elicitation is a type of study that aims at mapping inputs to emerging technologies through participatory design. The elicited inputs should be discoverable to novice users of systems [61]. A second product of elicitation studies is a better understanding of user behavior. Elicitation studies have shown that upper-body gestures are preferred in whole-body gesture systems [43], and that gestures produced are impacted by the size of the object [51, 55]. Elicitation has seen use for many input domains such as multi-touch surfaces [10, 34], and mobile devices [53], to internet of things use [63].

Elicitation studies typically use a Wizard of Oz (WoZ) experiment design [60, 61]. WoZ experiment design can be used to remove the gulf of execution between the participant and the system by removing the systems input recognizer [61]. In a WoZ elicitation experiment, a participant is shown a command (referent) to execute such as *move down*. The participant generates an input proposal for that referent which causes an experimenter to emulate the recognition of that input. In this work that is changed slightly to allow for better collection of speech results. For the command *move down* in this experiment, a participant was shown a virtual object moving down after which they would be asked to generate a command to produce that effect. By running the study this way we were able to collect inputs for a system that does not have a perfect recognizer or fusion model.

One outcome of an elicitation study is the production of a mapped set of inputs called a consensus set [10, 14]. More useful than a single set of mapped inputs is the observational data that comes from elicitation studies. This includes insight on the formation of inputs, the times surrounding input generation, and trends in user preferences for inputs and input modalities. An example of these extended benefits is the finding that the size of a gesture proposed is impacted by the size of the object shown [55]. This work extends previous gesture elicitation studies in AR [50] by testing the additional modalities of speech alone and multimodal gesture and speech interactions and allowing unconstrained gesture proposals for each referent. Furthermore, the set of interactions presented here shows the top few proposals allowing better interpretation of trends in gesture formation.

2.2 Gesture and Speech studies

A large portion of multimodal gesture and speech input studies have been focused on finding ways to combine them using multimodal fusion models [5, 11, 23, 47]. There has also been work on finding the timing windows for co-occurring gesture and speech interactions [30]. Some of this work looks at the usability of constrained sets of inputs such as limited gesture sets [13] or limited speech dictionaries [30]. These types of works look for a better understanding of a combination of the feasibility of inputs, the adaptability of people to constrained inputs, and the implementation or accuracy of fusion models for gesture and speech recognition. These works typically start with live mapped inputs and test usability or accuracy. **The work presented here is very different in that there are no constraints imposed on input proposals, and deliberate efforts were made to remove text based priming in the speech condition**. Participants are invited to generate any input proposal they see fit for the given referent and input modality.

While a few studies look at gesture and speech inputs have examined mid-air gestures [2, 11, 20, 27, 30, 37], some only looked at a subset of gesturing such as pointing gestures [7, 52], paddling gestures [21], or two dimensional (2D) gestures [35, 52]. The work presented here examines any mid-air gesture and / or utterance that a participant feels is appropriate for a given referent.

This study extends previous works done on multimodal gesture and speech elicitation [27, 37]. This extension is seen in the results reported and the methodology used. A previous study on interactions for computer-aided design program usage on 2d screens tested both gesture and gesture or speech interactions [27]. In that experiment, gestures were tested independently then gesture with optional speech was tested. This is different from our choice to examine each input individually. In both studies the referents were shown as animations, however, in this study participants were told that they were interacting with a system whereas Khan et al. asked participants to describe the referents to another person via a video chat [27]. The use case of computer-aided design as well as the choice of observing interactions compared to referent descriptions is markedly different, with examples of the referents used there being *extrude surface* or *pan*.

This work also extends the results of a study done on eliciting commands for television-based web browsing [37]. That study used paired elicitation where participants would sit in groups on a couch and propose either gesture, speech, or gesture and speech commands, as compared to the individual elicitation technique used here. That study also only examined the input modalities in a single pass where participants were allowed to produce any command in any modality or a combination of modalities. An important distinction is that referents were shown as text and read aloud by the experimenter in Morris, 2012 [37]. In this study we examine interaction proposals without text prompting.

This work differs from previous gesture and speech elicitation studies in several important ways. This work does not present users with any text when showing referents. Participants are not paired and are asked to produce an input for each modality. This is in comparison to prior works which commonly allows users to chose which modality they use when generating input proposals [59]. This work aims on finding intuitive inputs across the gesture, speech, and co-occurring gesture and speech interactions. This work does not attempt to improve gesture or speech recognition, nor does it attempt to build better multimodal fusion models. It is our hope that these results can be used towards those goals in future studies.

3 METHODS

3.1 Pilot Studies

Two versions of this study were run to assess the impact of referent display on proposal generation. The results of these pilot studies were used to inform the methodology decisions made in this experiment. These each used 6 people. In one of the pilot studies, we display the referents as text on the screen, which is different from our final design. The first pilot study's design is comparable to [37, 61]. In the second pilot study, we displayed the referent by showing the participants an animation of the intended effect of the interaction they would propose. The second pilot study's design is comparable to [27]. Both the pilot studies and this study tested the same input modalities, those being, gesture and speech, speech alone, and gesture alone.

In the first pilot study, there was evidence that text referents primed speech production. If the referent was *move right* the utterance was commonly "move right". This effect was more pronounced for translations, rotations had more variance in proposals but still showed signs of biasing. Repeating referents when producing speech proposals, such as saying "new tab" for the referent *new tab*, can be seen in the results of Morris, 2012 [37]. When the referents were shown as animations in the second pilot study, people would often mirror that animation in the gesture they produced. These mirrored gestures were often direct manipulations which are not uncommon in gesture interfaces [8], however, when designing inputs that priming could be problematic. The effect animations biasing gestures can be seen in the study done by Khan et al. 2019 [27], such as a *pan* gesture that mirrors the motion of the animation used.

This study's goal was to understand user speech behavior both alone and when co-occurring with gestures. With that in mind, we have chosen to show the referents as an animation. The only text shown to the participant was the input modality requested (e.g. "gesture only", "speech only", "gesture and speech"). This will allow us to have more robust speech results than when showing a text based referent. Another choice in elicitation methodology used in this experiment was to not have think aloud protocol as seen in [61]. The process of thinking out loud while generating speech proposals would confound the results, making speech data less reliable.

3.2 Methodology

This study was run as a within-subjects (i.e. repeated measures) elicitation study. The goal of this work was to gain a better understanding of the production of gestures, speech, and co-occurring gestures and speech when interacting with three-dimensional (3D) objects in an optical see-through AR-HMD. Participants were asked to generate proposals for gesture alone, speech alone, and multimodal gesture and speech interactions. These input modalities were presented in a counterbalanced order. Within each input, participants were asked to generate an interaction proposal for each referent. Meaning that a participant may be assigned the speech input modality first, then be asked to generate a speech proposal for each referent before progressing to either the gesture or gesture and speech condition. Referents were displayed in random order with each occurring once per input modality. The experimental setup is illustrated in Figure 1. Participants were told that they were guessing the interaction that someone in a different room was using to execute the referent they were presented with. A single referent sequence was a blank screen, a cube appearing, a 2-second pause, the cube playing an animation of the referent, then the participant proposing their input. The animation playing first removes the notion that the participant is directly interacting with the system. However, their belief that someone else is interacting with this system in a separate room, and the onscreen gesture aids (described later), caused the user to feel that this was a live system.

The referents (i.e. actions) that were used included the canonical manipulations (i.e. selection, rotation, positioning) found in [8] and the interactions that would be commonly used in a 3D manipulation or building task. They include translation and rotation on each axis, scaling, selection, and the creation or deletion of an object. This study looks at the use case of a 3D environment such as an architecture application, where objects must be manipulated and placed within that environment. This can be extended into interactive learning environments or data visualization environments where manipulating virtual content can provide better learning outcomes [48]. Most

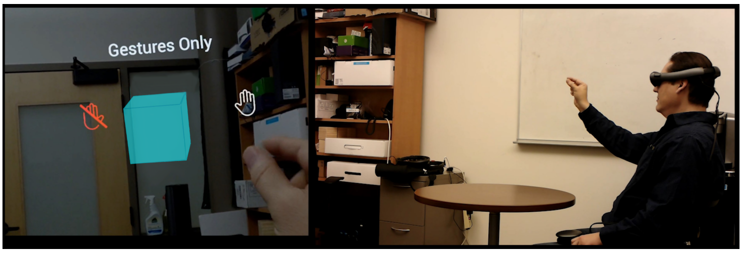

Fig. 1. Experimental Set up: Left, participant view, Right: participant

optical see-through AR-HMDs (e.g., Magic Leap One) and some VR-HMDs (e.g., Oculus Quest) have built in ego-centric sensors. With that in mind, the gestures in this study were analyzed by viewing the ego-centric interactions within the environment.

The metrics used for gesture proposal interpretation are Agreement Rate ($\mathcal{AR}$)[1], co-agreement rate ($\mathcal{CAR}$), and the (V_{rd}) significance test [57, 58, 61]. $\mathcal{AR}$ is the proportion of proposals in agreement over the total possible proposals pairs in agreement. High $\mathcal{AR}$ can be interpreted as more consensus among participants in the proposals generated for a given referent. This metric is used at the referent level meaning that a given proposal will not have an associated $\mathcal{AR}$ but a referent will. Based on distributions of $\mathcal{AR}$ over various sample sizes participants an $\mathcal{AR}$ of 0.3 has been said to indicate high agreement given our N of 24 [57]. The V_{rd} is a test of the difference in agreement rates between k referents. A low p-value indicates that there is a difference between the tested referents. The $\mathcal{CAR}$ can be seen as the percent of participants that agree on proposals for k referents. Fleiss' Kappa and the associated chance agreement term are used to justify using an $\mathcal{AR}$ of 0.3 as high [56].

For speech proposal analysis the consensus-distinct ratio ($\mathcal{CDR}$) and max-consensus ($\mathcal{MC}$) were used [37]. The $\mathcal{CDR}$ is the percent of matching proposals that have been suggested by more than a recommended baseline of two participants out of all the proposals for a given referent [37]. $\mathcal{MC}$ is equal to the percent of participants proposing the top-ranking proposal. The combination of these metrics can be used to see the peak and spread of speech proposals.

3.3 Participants

The study consisted of 24 volunteers (10 Female, 14 Male). Participants were recruited using emails and word of mouth. Participants were 18 - 46 years old (Mean = 25, SD = 6.9). Six participants had less than half an hour of previous AR-HMD usage experience, the other participants had no prior device usage. All participants reported normal or corrected to normal vision. Five participants reported being left-handed. Five participants reported weekly use of VR. Only one of 2 of those participants used VR more than 5 hours weekly (5 hours, 10 hours), the rest were 1-3 hours weekly.

3.4 Procedure

For each session participants started by completing the informed consent and demographic questionnaire. That questionnaire asked about prior device usage (AR, VR, multi-touch), age, handedness, vision, and gender. A two-minute instruction video was shown describing the experiment after which the participant could ask the experimenter questions. During the video, they were told that any utterance or gesture either one-handed or two-handed, produced was acceptable. The participant would then don the AR-HMD and complete a practice trial for each input modality.

[1] Please note that agreement rate $\mathcal{AR}$ uses a different font to avoid confusion with AR for augmented reality.

During the practice trials, the participant could ask any questions they had and adjust the device. Participants were also alerted to the devices gesture recognition aid shown (Figure 1) during the practice. This aid was an image of the outline of a left and right hand. The hands were white when a participant's corresponding hand was inside the device's gesture sensing range. They would be red with a line through them when the participant's corresponding hand was outside of the device's recognition range. This aid was provided to help prompt participants to generate gesture proposals that could be used in AR-HMDs as well as to add more immersion to the interactions with the object in the experiment. As this was a WoZ study, the aid was only adding realism to the task, no gestures were recognized.

The referents were shown as animations (showing the object then moving it left over 2 seconds for the referent *move left*). No text was shown to the user. For three referents animations that were not basic movements had to be shown. For the *create* and *delete* referents particle effects of an object appearing or disappearing over two seconds were used. For the *select* referent, the object was highlighted by increasing its hue and adding a light outline. Each referent was presented as a cube rendered 50cm in front of a user's display. The modality to use for the proposals was shown as text above the cube. The experimenter would trigger the loading of the next referent a few seconds after a proposal was generated by the participant. The new referent would always appear in the center of the participant's display, stay there for 2 seconds, then execute the animation for the referent.

3.5 Apparatus

This experiment was conducted using a Magic Leap One optical see-through AR-HMD. The WoZ system was developed in Unreal Engine 4.23.0. A Windows 10 professional computer with an Intel i9-9900k 3.6GHz processor and an Nvidia RTX 2080Ti graphics card was used for development. Data were recorded on the Magic Leap One. A GoPro hero 7 black was used to record an ego-centric view of the interactions for analysis. A 4k camera was used to record an exo-centric view of the interactions as a backup to the GoPro.

4 RESULTS

4.1 Gestures Proposals

4.1.1 Gestures from the unimodal gesture block. The average $\mathcal{AR}$ observed for the gesture block was 0.302 with $\kappa_F = .257$. Given our sample size of 24 and the low chance agreement term ($p_e = .052$) used in Fleiss' Kappa coefficient we consider rates above 0.3 as high levels of consensus [56, 57]. Agreement rates are shown in Table 1.

Table 1. Agreement rates per referent by block

	Create	Delete	Enlarge	Move Away	Move Down	Move Left	Move Right	Move Towards	Move Up	Pitch Down	Pitch Up	Roll C	Roll CC	Select	Shrink	Yaw Left	Yaw Right
Gesture	0.21	0.11	0.28	0.37	0.38	0.49	0.44	0.28	0.49	0.16	0.28	0.56	0.39	0.09	0.14	0.25	0.22
Gesture and Speech	0.09	0.05	0.18	0.50	0.29	0.33	0.32	0.34	0.35	0.19	0.22	0.28	0.45	0.09	0.14	0.15	0.25

Legend: C: clockwise, CC: counterclockwise, Highlighted cells have high agreement

The effect of referent type on agreement rates was observed to be significant ($V_{rd(16,N=408)} = 510.342, p = .001$). High agreement was found for each of the translation referents except *move away*, and for both the *roll clockwise* and *roll counterclockwise* referents (Table 1). The highest

$\mathcal{AR}$ was found in the *roll clockwise* referent ($\mathcal{AR}_{roll\ clockwise} = .56$). A mapping of the frequency of gesture proposals with more than three participants suggesting them and the corresponding referents can be seen in Figure 2. The gestures from the gesture block have "G" next to the referent name.

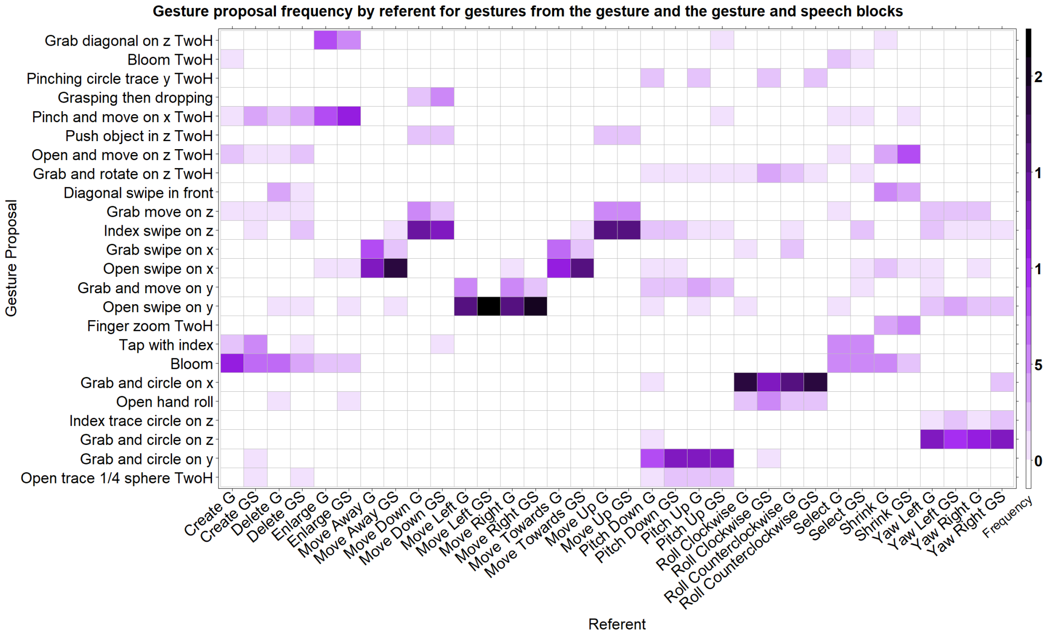

Fig. 2. Gesture proposal frequency by referent for gestures from the gesture and the gesture and speech blocks

Legend: G: Gesture Block, GS: Gesture and Speech block, TwoH: Two handed gesture, Open: fingers open, Grab: hand closed, Pinch: two or three finger pinching, z: up, x: forward, y: side

The more abstract referents, *Create*, *Delete*, and even *select* exhibited low agreement rates ($\mathcal{AR}_{shrink} = .14$, $\mathcal{AR}_{delete} = .11$, $\mathcal{AR}_{Select} = .09$). This is mostly due to disagreement between proposals shown by an increase in the count of colored cells in Figure 2. Common hand poses and movements are shown in Figure 3. *Select* had low $\mathcal{AR}$ due to participants having a difficult time interpreting the referent animation. *select*'s animation showed the cube normally (left side of Figure 1) then gradually becoming highlighted by reducing the hue after a 2-second delay. In pilot tests on the *select* referent we attempted other visualizations such as bouncing, or an arrow appearing and pointing at the cube. These animations primed the speech and gesture produced. The highlight animation had the highest rate of participants guess what it was, but that rate was still fairly low.

The translation referents (*up, down, left, right, away,* and *move away*) had high gesture agreement among participants ($\mathcal{AR}_{translations} = .432$). Among these translational referents, the direction of motion displayed a significant effect on agreement rates ($V_{rd(5,N=144)} = 52.765, p < .001$). A significant difference in agreement was observed for referents *towards* and *away* ($V_{rd(1,N=48)} = 9.921, p < .01$). *Roll clockwise* and *roll counterclockwise* had high $\mathcal{AR}$with an average ($\mathcal{AR}_{roll} = .475$). This was higher than the average $\mathcal{AR}$ for all the rotational referents ($\mathcal{AR}_{rotations} = .31$) which drops to ($\mathcal{AR}_{rotations\ without\ roll} = .23$) when roll is removed. We believe that participants may not have had much experience with altering the pitch or yaw of virtual objects and this is reflected with the low $\mathcal{AR}$. The excepting being roll manipulations, which seem more common with objects like clock hands moving that way, inflating their $\mathcal{AR}$.

There was low $\mathcal{AR}$for *shrink* and *expand*, which is surprising due to the prevalence of touchscreen phones and near-daily use of the two-finger zoom-in and zoom-out commands. Those gestures occurred with some frequency, however, there were a high number of two-handed comparable gestures proposed (Figure 2). For these people would pinch either corner and pull or push their hands away or towards other either diagonally or horizontally.

The heatmap in Figure 2 helps show the trends among gesture proposals, darker colors indicate more proposals. The gestures mapped are all reversible gestures meaning a movement in the opposite direction is the mirror of the gesture. An example of this is seen in the gesture for *move up* which was a palm up push up where *move down* was a palm down push down. The referents *move left* and *move right* had very few different proposals indicating high agreement on the appropriate gesture. Whereas, referents like *select* had a high range of proposals given. When examining the plot horizontally by proposal instead of vertically by referent trends in how participants map the same gesture to multiple actions are seen. For example, an open hand swipe either left or right was used for 9 referents. The uses make sense, a quick swipe from right to left could be seen as deleting an object, or touching the side of an object and moving left or right would change its yaw. The "Bloom" gesture was used for every abstract referent. The variations present in some manipulations were only in the pose of the hand, or the number of hands, but not the motion of the gesture. *Move up* had three common proposals with each centering around some sort of grab and a movement on the z-axis.

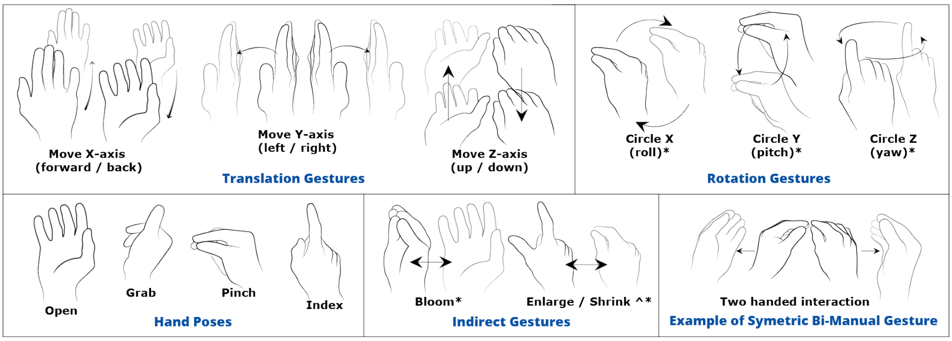

Fig. 3. Hand pose examples, two handed gesture example, and common gestures by category of movement or type of gesture

Legend: *: reversible gesture, $\wedge$: commonly two handed, z: up/down, x: forward/back, y: left/right

4.1.2 Gestures from the multimodal gestures and speech block. The results for the gesture proposals from the gesture and speech are very similar to the gestures from the gesture alone block. By comparing columns with the matching referent names (e.g. *create G* and *create GS*), an image of the differences of proposals across these blocks can be drawn. The overall agreement rate observed for the gestures in the gesture and speech block was 0.247 with $\kappa_F = .218$. The low chance agreement term ($p_e = .037$) used in Fleiss' Kappa coefficient indicates an agreement beyond chance [56], allowing us to consider $\mathcal{AR}$ rates above 0.3 as high [57]. The agreement rates for each referent are shown in Table 1.

The effect of referent type on agreement rates was observed to be significant ($V_{rd(16,N=408)} = 904.091, p = .001$). High agreement was found for each of the translation referents except *Move Down* $\mathcal{AR}_{Move\ Down} = .29$. This was caused by an increase in the number of “drop” gesture proposals. *Roll counterclockwise* also exhibited high $\mathcal{AR}$ ($\mathcal{AR}_{Roll\ counterclockwise} = .45$) (Figure 1). The highest $\mathcal{AR}$ was found in the *Move Away* referent ($\mathcal{AR}_{Move\ Away} = .5$). A mapping of the frequency of the top gesture proposals and the corresponding referents can be seen in Figure 2.

The abstract referents *Create*, *Delete*, and *select* exhibited low agreement rates ($\mathcal{AR}_{Create} = .09$, $\mathcal{AR}_{Delete} = .05$, $\mathcal{AR}_{Select} = .09$). This is mostly due to disagreement between proposals shown by an increase in the count of colored cells in Figure 2. As in the gesture block, *select* had low $\mathcal{AR}$ due to participants having difficulties interpreting the referent’s animation. The translation referents (*up, down, left, right, away*, and *move away*) had high gesture agreement (average $\mathcal{AR}$ = .355). A significant disparity was observed for referents *roll clockwise* and *roll counterclockwise* ($V_{rd(1,N=48)} = 59.522, p = .001$). *Roll clockwise* and *roll counterclockwise* had high $\mathcal{AR}$ with an average of ($\mathcal{AR}_{Roll} = .475$. This was higher than the average $\mathcal{AR}$ for all the rotational referents ($\mathcal{AR}_{Rotations} = .31$) which drops to ($\mathcal{AR}_{Rotations\ without\ roll} = .23$) when roll is removed. We believe that participants may not have had much experience with altering the pitch or yaw of virtual objects and this is reflected with the low $\mathcal{AR}$. As in the gesture block the scale referents had low $\mathcal{AR}_{Shrink,Enlarge} = .18, .14$.

The bulk of the gestures shown in Figure 2 are direct manipulation gestures. Translations are concentrated in a few gestures where rotations are spread across more proposals. Even so, most rotation proposals involved tracing or moving a participant’s hand in a circle. In the case of most of the referents, there was an increased spread of gesture proposals in the gesture and speech block. This was not the case for every referent, some such as *move left* and *roll counterclockwise* have a decreased number of proposals in the gesture and speech block. Largely the gestures used did not change drastically between the two blocks.

4.2 Speech Proposals

Displaying the referent in elicitation studies [42] and reading the referent aloud in gesture and speech elicitation studies [37] both have precedence. These practices can prime the utterances proposed. When interpreting these results remember that neither think out-loud protocol nor text was used for referents. The participant only saw an animation of the referent being executed. When analyzing speech proposals we have dropped the object specifier to remove a level of increased proposal complexity. We believe that if an object is already selected, using the command "Move the cube right" and "move right" could be reasonably considered the same, the exception being the *select* referent.

Table 2. Frequency of syntax format by block

	<action>	<action> <direction>	<action> <object> <direction>	<action> <object>	<direction>
Speech	28.19%	47.06%	14.22%	9.31%	1.23%
Gesture and speech	38.48%	39.95%	12.99%	6.86%	1.72%

4.2.1 *Speech from the unimodal speech block.*

While were told that any utterance or sentence was acceptable, they primarily stuck to <action> <direction> or <action> <direction> syntax structure. The rates for syntax are found in table 2. The difference between <action> <direction> and <action> <object> <direction> was only a descriptive specifier of the object (e.g. “cube”). The

<action> and <direction> words were the same as found when no specifier was used (e.g. “move the cube left” would be “move left”).

The $\mathcal{MC}$ and $\mathcal{CDR}$ for this block are shown in Figure 3. Note that $\mathcal{MC}$ is equal to the percentage of participants proposing the top proposal per referent, shown in the "Top proposal" column in Table 3. *Yaw* referents had some of the highest $\mathcal{CDR}$ indicating high disagreement among participants on the utterances proposed ($\mathcal{CDR}_{\textit{Yaw left,Yaw right}} = .62, .78$). *Delete* also had a high amount of disagreement among proposals ($\mathcal{CDR}_{Delete}$ = .57). Both *create* and *shrink* had low $\mathcal{CDR}$ ($\mathcal{CDR}_{Create,\,Shrink}$ = .18, .25). Low $\mathcal{CDR}$ means that most participants grouped around the top proposals. The rest of the referents hold moderate disagreement values.

The highest $\mathcal{MC}$ value belongs to *move up* ($\mathcal{MC}_{move\ up}$ = .54). Most participants proposed either "Move up" (54.17%) or "go up" (12.5%). The full list of each referent’s top two proposals and the percent of participants proposing them can be seen in Table 3. For the translational referents "move" was used as the <action> command in either the top or second place proposal. *Move down* ($\mathcal{MC}_{move\ down}$ = 33.33%), which had “drop” as the top proposal, was the only translational referent that did not have “move” in it. The second-place proposal for *move down* was “move down” (29.17% proposed). The referents for *move up*, *left*, and *right* all had the directional term (up, left, right, down) included. *Move towards* and *move away* had either towards, and forward, or away, and back proposed as the <direction> term. This indicates that aliasing “away” with “back”, and “towards” with “forward”. Aliasing commands has been suggested as being beneficial when dealing with unimodal speech [37, 61]. Note that these terms are reversible, which was a common trend with most opposite proposals (e.g. “appear”, “disappear”).

For the rotational referents (*pitch*, *roll*, *yaw*) the average $\mathcal{MC}$ was 24.31% which is lower than the translations average $\mathcal{MC}$ of 35.42. For each rotation the action was specified by either “spin” or “rotate” in all of the top proposals by participants (Table 3). This is not unexpected, the terms “roll”, “pitch”, and “yaw” are uncommon in most fields. *Pitch* has the most unique mapping of proposals commonly “towards”, “away” for *pitch up* and “back” for *pitch down. Roll* and *yaw* have the terms “left” and “right” for directions. We believe that this ambiguity is solved by adding gestures to indicate the “spin” direction, or by an expert assigning speech commands such as “spin clockwise” in the *roll clockwise.*

The referents *create* and *delete* had single word commands for the top and second place proposals as well as some of the higher $\mathcal{MC}$ found ($\mathcal{MC}_{create,delete}$ = 41.67%, 50%). The top proposals were “appear” and “disppear”. These proposals could be considered similar to the reversible gestures found in this study and others [50, 61]. “Create” appeared as a second place proposal (20.83%) and “delete” was a third place proposal (12.5%). *Shrink* ($\mathcal{MC}_{shrink}$ = 45.83%) also had a high agreement between participants. As did *enlarge* ($\mathcal{MC}_{enlarge}$ = 37.5%). *Select*, with its difficulties in animating had low agreement and high spread of proposals ($\mathcal{CDR}, \mathcal{MC}_{select}$ = .55, 21%).

4.2.2 Speech from the multimodal gesture and speech block. A chi-square test of independence showed that there was a significant association between the block and syntax choice ($X^2(4, N = 408) = 10.928, p < 0.03$). Participants used a higher rate of <action> only syntax than found in unimodal speech. <Action> <direction> syntax use was reduced by 7.11%. The rates for the syntax are found in Table 2. Both of the syntax structures that used an <object> specifier were lower in this block. Most often when an object would have been specified it was replaced by a gesture indicating the object. This gesture was often reaching out and grabbing or another type of direct manipulation.

The average $\mathcal{MC}$ for the translational referents decreased by 10.33% from the speech block (Figure 3). This was due to more participants using the <action> syntax. The $\mathcal{CDR}$ did increase in the translational referents as well. Participants had less agreement on the appropriate proposal and

Table 3. Speech proposals for the speech from the speech block and the speech from the gesture and speech block

	Speech from the speech block					Speech from the gesture and speech block				
Referent	Top proposal	$\mathcal{MC}$	2nd place	$\mathcal{MC}$	$\mathcal{CDR}$	Top proposal	$\mathcal{MC}$	2nd place	$\mathcal{MC}$	$\mathcal{CDR}$
Create	appear	41.67%	create	20.83%	0.18	appear	33.33%	create	29.17%	0.18
Delete	disappear	50%	remove	16.67%	0.57	disappear	54.17%	make disappear	12.5%	0.33
Enlarge	enlarge	37.5%	grow	16.67%	0.36	enlarge	25%	grow	20.83%	0.56
Move Away	move back	25%	move away	12.5%	0.38	move back	16.67%	push away	16.67%	0.64
Move Down	drop	33.33%	move down	29.17%	0.44	drop	29.17%	move down	16.67%	0.46
Move Left	move left	37.5%	slide left	20.83%	0.44	move left	25%	slide left	16.67%	0.2
Move Right	move right	41.67%	slide right	20.83%	0.44	move right	20.83%	slide right	20.83%	0.33
Move Towards	move forward	20.83%	move towards	12.5%	0.36	move forward	16.67%	move towards	12.5%	0.43
Move Up	move up	54.17%	go up	12.5%	0.33	move up	41.67%	go up	8.33%	0.33
Pitch Down	rotate	20.83%	rotate towards	16.67%	0.46	spin forward	20.83%	rotate towards	16.67%	0.6
Pitch Up	rotate away	16.67%	spin backward	12.5%	0.5	spin back	16.67%	rotate	12.5%	0.43
Roll C	spin right	20.83%	rotate	16.67%	0.5	rotate	20.83%	rotate right	16.67%	0.36
Roll CC	spin left	25%	rotate left	20.83%	0.4	spin left	25%	rotate	16.67%	0.23
Select	glow	20.83%	highlight	20.83%	0.55	change	25%	glow	25%	0.36
Shrink	shrink	45.83%	minimize	8.33%	0.25	shrink	41.67%	make smaller	8.33%	0.23
Yaw Left	spin left	33.33%	rotate	16.67%	0.62	spin	29.17%	rotate left	16.67%	0.36
Yaw Right	spin right	29.17%	rotate	12.5%	0.78	rotate right	20.83%	spin	20.83%	0.6

Legend: C: Clockwise, CC: Counterclockwise, $\mathcal{MC}$: Max-Consensus, $\mathcal{CDR}$: Consensus-Distinct Ratio

the spread of proposals was wider. Even with the differences in syntax use between blocks, the top choice proposals remained the same.

The rotational average $\mathcal{MC}$ only decreased by 2%, the $\mathcal{CDR}$ decreased by 0.113. This means that while agreement on the top choice proposal was negligibly impacted between blocks, the spread of proposals given in the gesture and speech block for rotations was narrower than in the speech block. Most of the top choice proposals for translations changed between the two blocks (Table 3). Some switched from using "spin" to "rotate" or vice versa. As an example, the proposal for *yaw right* switched from "spin" to "rotate" while the top proposal for *roll clockwise* did the opposite. We take this to mean that the words "rotate" and "spin" are without a clear mapping in participants' minds. For translations gesturing removes much of the ambiguity by allowing for a physical motion to indicate the intended rotation direction.

Most proposals remained the same between the two blocks with slightly different $\mathcal{MC}$ rates. There was a shift in *create* from the top choice proposal of "appear" from ($\mathcal{MC}_{Create} = 41.67$) to ($\mathcal{MC}_{create} = 33.33$) in the gesture and speech block. This is captured in an increase of 8.34% in the second choice proposal in the gesture and speech block. *Delete* was mostly unchanged in top proposals but did have a decreased $\mathcal{CDR}$ ($\mathcal{CDR}_{Delete} = .33$). Meaning there were less distinct proposals made. *Enlarge* had a lower $\mathcal{MC}$ and higher $\mathcal{CDR}$ in the gesture and speech block ($\mathcal{MC}, \mathcal{CDR}_{enlarge}$ = 37.5%, .56).

4.3 Co-occurring gestures and speech proposals

When looking at pairings of speech and gesture proposals in the gesture and speech block the agreement rates fall drastically due to the highly nuanced nature of speech. Individually each modality had referents that experienced high levels of agreement. For gestures refer to Figure 2 and Table 1. For speech consensus refer to Table 3. We feel that matching common syntax structure with gestures when looking at multimodal gesture and speech interactions is more beneficial than observing the pairing of utterances with gesture proposals. The speech syntax by block is shown in Table 2. Gesturing remains consistent in both conditions indicated by a high p-value in a chi-square test ($X^2(49, N = 408) = 10.928, p < 0.247$) (Comparing G and GS in Figure 2). The same is true of speech (Compare S and GS in Table3). This is beneficial in a few ways. In the case of translations and scaling it allows each input to serve as a back up to the other. Allowing for mutual disambiguation as found by [45]. In the case of rotations, the gesture provides context on the direction of rotation while the speech was commonly “spin” and a direction. With abstract commands, the same gesture, a “bloom” gesture, was found for multiple referents. In those cases, speech allows interpretation of which command is being executed with the gesture.

4.3.1 Timing of co-occurring gestures and speech. In the gesture and speech block the time windows of phases of a co-occurring gesture and speech interaction were measured based on the time of gesture initiation. These were collected from videos of the experiment and hand-annotated. The phases used to describe interactions are gesture initiation, stroke start, speech start, stroke stop, and speech stop. These are taken from McNeil’s segmentation of co-occurring gesture and speech interactions [33]. The gesture start is the first perceptible movement made by someone. Speech start is the first perceptible sound being made. For both of those if a false start was found it was discarded and the time of the next movement was taken. As an example, if a participant said “Ummm” then later said “move”, the time of “move” was used. A stroke is considered to be the segment of a gesture that holds the information content of the gesture, as well as the peak of effort in that gesture [33]. Gesture stroke was found by measuring the time of the first visible change in the direction of the gesture. The stroke stop was the last change in direction and was found by reversing from the end of a gesture. A full gesture interaction would look like someone starting to move their hand in preparation for a stroke (gesture start), starting a meaningful gesture (stroke start), then ending the gesture (stroke stop). The hand moves up in preparation, pushing the object forward, then retracts to its initial state.

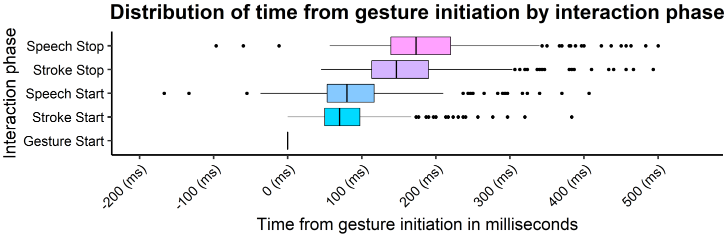

Fig. 4. Distribution of time from gesture initiation by interaction phase

Shapiro-Wilks tests show that the time information took a non-normal distribution of each of the phases at ($p < .001$). Bonferroni adjusted Wilcoxon rank-sum pairwise comparisons indicate that each phase's time is significantly different from each other. The p-values were ($p < 0.001$) in each comparison except between "stroke start" and "speech start" which was ($p = 0.03$). The descriptive statistics for times by the phase of interaction are shown in Table 4.

We find that in this experiment speech nearly always occurs after a gesture is started (Figure 4). The difference in start time is around 81.667 ms. Importantly, the information content of the gesture, the stroke, starts commonly 90.872 ms after gesture start (Table 4). This means that by watching a gesture's changes in direction, we can predict when speech will occur, and when a meaningful message is communicated. Strokes were found to end before speech 23.739 ms. The total interaction from start to finish was typically 187.566 ms. Most speech proposals were only 2 words so this relatively short interaction time makes sense.

Table 4. Time from gesture start for phases of an interaction in milliseconds

	Gesture Start	Stroke Start	Speech Start	Stroke Stop	Speech Stop
Mean	0	90.872	81.667	163.827	187.566
Standard Deviation	0	70.548	54.743	78.742	80.064
Standard Error	0	3.493	2.710	3.898	3.964

These results are similar to previous work [7, 30], though slightly quicker and more granular. These results expand time windows from being formed for pointing gestures only [7], and show that these time windows follow similar patterns for deictic and manipulative gestures. They also show that gesture and speech interactions in AR-HMDS have similar timings [32] and patterns of occurrence [54] as in other environments.

5 DISCUSSION

The hand positions found here were similar to the ones observed by Piumsomboon et al. [50]. The gesture proposals were commonly single-handed. This is similar to findings on multi-touch surfaces [28, 38, 39] and mid-air full-body studies [42]. For manipulations users often interacted based on that actions real-world corollary. This is evident in the translation gestures which were predominantly some form of directly pushing the surface of the object. This theme of interaction was seen with manipulation gestures in previous work [50]. We speculate that the similarities in proposals were due to the object being rendered in the participant's real-world view by use of optical see-through AR. With that, users would interact based on their interpretations of naïve physics when possible [22]. This was mostly true for rotations which were accomplished by either grabbing some part of the object and moving their hand in circle motions as also seen in Piumsomboon et al.'s study [50]. The exception to these similarities is in the occurrence of "index extended" circular motions as an indirect gesture.

Scaling was often a two-handed pinch and drag gesture which was more common than touch screen "zoom in" and "zoom out" gestures. Grabbing the corners or sides of an object would correspond with how a mental model of a stretchable object would be manipulated. This gesture was seen for scaling on an axis in [50]. Similarities in gesture proposals between these studies start disappearing as the referents become more abstract. This can be seen when comparing the proposals for *delete* which was a "grasping" gesture in other work [50] and a "bloom" gesture here.

That most of these gesture proposals extend across two studies and two-time points is a strong indication that these gestures and hand poses should be candidates for inclusion in future AR interaction systems. This study did not ask participants to reserve proposals for a single interaction (i.e., a bloom could be used for *create* and for *select*). Redundantly mapped proposals showed up more in the abstract referents. In the work of Piumsomboon et al. participants were asked to refrain from redundantly mapping inputs [50]. The similarities of proposals between these works show that requiring unique interactions may not have greatly impacted many of the gesture proposals [50]. An interesting, redundantly mapped gesture was the “index swipe” which was used for both *yaw* and *move up/down.*

We feel that the combination of high levels of agreement for translations in the gesture block and the tendency to have more unique proposals given in the gesture and speech block indicate that unimodal gestures are well suited for object manipulations. While rotations had a high number of single-hand “grab and rotate” gestures, many were indirect manipulations using a index finger and tracing a circle. For these, a non-isomorphic gesture seem well suited. The most agreed-upon proposals for manipulations were all reversible gestures. Indicating a preference for reversible gestures which mirrors previous work [50, 61].

Some of these direct manipulations were implemented and tested against a gesture+speech interface in the work of Piumsomboon et al. [49]. The findings were similar to the user stated expectations observed here. When specific degrees or units were needed participants indicated a preference for speech. For most basic or single object manipulations, gesture seemed preferred across both studies [49]. Peoples’ preference for multimodal interactions typically increases as a task’s cognitive load increases [46] or the task’s complexity increases [49]. We expect that if more complex referents were used the user stated preference for multimodal interactions would have been higher.

Gestures showed less usability for the *create* and *delete* referents. Speech had more clarity in these cases with common utterance being “appear” and “disappear”. Gesture proposals for abstract referents were consistently the “bloom” gesture, which was proposed for many referents, and thus hard to interpret without additional context. Speech show more promise for use with abstract commands and conceptually difficult actions that do not map well to a user’s mental model. An example would be opening a new browser window, which was not tested here. Speech proposals for both *create* and *delete* had high agreement, emphasizing this strength.

When used together gestures and speech provide different benefits based on the type of referent being executed. For translations and scaling this was commonly redundancy, which allows for error correction in a recognizer system. For rotations, this pairing allows a clear communication of the desire to rotate then clarifying the direction with a co-occurring gesture. This allows for intuitive interactions with mutual disambiguation with information from the complementary channel. An added benefit of allowing speech and gesture for rotations is the ability for participants to communicate the degrees of rotation, allowing for more accurate interactions.

In the speech condition participants preferred to use <action> <direction> or <action> <object> <direction> syntax over complete sentences. Implying that both unimodal and multimodal speech utterances are syntactically simplified compared to conversational speech [44]. This is seen as saying “move” and “finger flicking” in the direction of the intended movement. In either case, the intended <action> was present indicating that full natural language processing may not be necessary for basic multimodal interactions.

This work contributes to findings on multimodal interactions and touches on some of the potential pitfalls of referent display which would cause reproduction to be difficult, as mentioned by Villarreal-Narvaez et al. [59]. The impact of referent display on proposals is seen most saliently in the low $\mathcal{AR}$ for the *select* referent which often received high $\mathcal{AR}$ in prior studies [43, 50]. The timing information and patterns here provide insight into the formation of these interactions and extends the timing windows constructed by Lee et al. [30] by adding the phase of the interaction by the time of that phases initiation. This study gathers proposals within each modality allowing for comparison against gesture only studies [50], while also contributing to the less common multimodal elicitation literature [27, 37].

6 DESIGN GUIDELINES

Instead of directly proposing a single set of consensus interactions within each modality we have chosen to show the distribution of interactions. By looking at these distributions a picture of trends across the top few proposals can be seen. For some referents, such as the translational referents, the top gesture in the gesture and the gesture and speech block matched (Figure 2). For translations often the top proposal was a reversible swiping gesture for moving the object in the x-axis and y-axis and an index extended swipe for movement on the z-axis. The speech proposals in these cases were also reversible (Figure 3). The first choice in all translations except *move down* was to say "move" and then a direction. For *move down* people commonly said "drop". *Create* and *destroy* followed the same pattern with a reversible bloom gesture either starting closed then opening or starting open then closing and the utterances "appear", and "disappear". For most gestures, a bi-manual version that was a symmetric two-handed version of the uni-manual proposal was also proposed (i.e. pushing with one open hand versus pushing with two).

Most gestures were based on the participants' understanding of naive physics, meaning how they perceived an object would react to an interaction as it would in the real world. Most variations occurred within specific hand poses but not the larger movements of the hand/arm. As such we recommend aliasing manipulative gestures across hand positions (open hand, pinch, grab) based on the type of movement. A second consideration should be made on the inclusion of bi-manual gestures. while not found in abundance here, other work [27, 50] has found evidence that users may gravitate towards using them in other domains and with larger objects [51, 55].

Other referents had less consistency. In the case of *shrink* and *enlarge* a "bloom" gesture and two handed "pinch and drag" gestures were common. In this case, we would suggest reserving the "bloom" gesture for *create / delete* and allowing "grab and pull" and scaling as seen both here and in earlier work [50]. The top speech proposals for scaling were more agreed upon and should be implemented as well. Those were the reversible pair "enlarge", and "shrink". Rotational referents other than *roll clockwise* have high levels of disagreement among proposals. "spin" and "flip" should be enabled as action selection words then a gesture should be allowed for controlling the direction of the rotation.

Direct manipulations should be allowed when possible, especially for basic manipulations. Speech and gesture as multimodal interactions showed promise in areas where one or the other input lacked and should be included. Implementing a system such that it has an internal model of functionality that aligns with what most participants formed as their mental model of functionality would increase the user's chances of guessing the inputs. This would be most easily achieved with direct manipulations, which in this study were often very close to their real-world corollary.

Participants seldom used full sentences or referred to the object being manipulated (Table 2). Due to that word spotting should be sufficient for most tasks. Only two participants used full sentences and those sentences followed the <action> <object> <direction> syntax with prepositional terms added (e.g. "move to the right" compared to "move right"). In either command, the actual information content is held in the <action> <direction> terms which could be spotted. The use of simple commands when possible was also observed by [30].

The windows built around co-occurring interactions are incredibly useful to systems needing to decipher interactions. With segmenting interactions based on the first movement of a gesture, the transition into the stroke phase, the information content of both the speech and the gesture portions of the interaction can be found. In this study gestures nearly always preceded speech (405/408 proposals). Most commonly speech was around 81.67 milliseconds after a gesture initiated. The stroke was often 90.87 milliseconds after the start of a gesture. Both of those phases represent the initiation of the actual information content of the interaction. The back end of these interactions is slightly less concrete. Often the end of a gesture preceded the end of an utterance. A system could be designed to use a time-out window after which the speech would be considered a separate interaction.

7 LIMITATIONS OF THE STUDY

By choosing to show animations for referents the gesture proposed may be biased to follow the animations shown. This choice was made to preserve the value of the speech proposals with pilot studies that showed speech was less impacted when showing the animations of the referents as opposed to the text. This study only allowed one proposal per referent per block. Having participants propose more than one interaction may have generated interactions that they felt more well suited to the referents. This study only showed a single virtual object at a time, which would impact the selection phase of any interaction. To help compensate for this we used the referent *select* independently.

For the rotational referents participants would sometimes use misaligned gestures and speech. They might say “roll clockwise” and perform a counterclockwise movement with their hand. Multimodal systems can suffer from compounding errors caused by incorrect recognition, or mismatched interactions such as the ones seen in this study [6]. These errors could take more time than standard uni-modal errors to correct or cause compounding errors when a second error is made during an attempt to correct the first.

8 CONCLUSIONS AND FUTURE WORK

Several questions remain unanswered. If there were more than one object shown the gesture results would show more selection gestures. The choice of an object used could also impact the production of interactions. If a larger object or a differently shaped object was presented the hand postures used may differ. Future work should involve testing the proposals found here against ones produced by text-based referents to assess the impact of referent display.

Compound errors in uni-modal text entry systems cause a generally linear increase in correction time [3]. Recent work has shown that improved error correction methods can reduce the time it takes users to reconcile text entry errors, decreasing the overall amount the user is slowed down by the error correction process [1]. Further work is needed to examine whether this holds true for multimodal interactions.

This work presents a within-subjects elicitation study across three input modalities (gestures, speech, and co-occurring gesture and speech). By examining each modality independently direct comparisons between the changes in speech and gesture from unimodal interactions to multimodal interactions are shown. Trends in gesture proposals are shown at a granular level. Highlighting that while there is often disagreement in proposals given, that disagreement manifests as variations in with similar underlying formations. In gestures, this was a variation of the hand position and not in the gross movement. In speech, this disagreement is seen as consistency in the <direction> phrases used and minor variations in the <action> phrase (e.g. “move” to “go”). While a singular mapping of the top proposals would yield a consensus set that is discoverable to most users, by aliasing and understanding the likely variations in interactions, a larger percentage of users’ natural interaction preferences can be captured.

This work extends the work of linguists [26, 33, 36], and the work of computer scientists [5, 7, 15, 31] into AR-HMD building environments. Timing windows based on the phases of co-occurring gesture and speech interactions as described by McNeil [33] have been constructed. Showing that in HCI the gesture stroke is closely aligned with the information content of both the gesture and the utterance given. These windows can be used to construct more accurate multimodal fusion models.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) awards NSF IIS-1948254, and NSF BCS-1928502 and the Defense Advanced Research Projects Agency (DARPA) ARO contract W911NF-15-1-0459.

REFERENCES

[1] Ohoud Alharbi, Ahmed Sabbir Arif, Wolfgang Stuerzlinger, Mark D. Dunlop, and Andreas Komninos. 2019. WiseType: A Tablet Keyboard with Color-Coded Visualization and Various Editing Options for Error Correction. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019* (Kingston, Canada) *(GI'19)*. Canadian Human-Computer Communications Society, Waterloo, CAN, Article 4, 10 pages. https://doi.org/10.20380/GI2019.04

[2] Dimitra Anastasiou, Cui Jian, and Desislava Zhekova. 2012. Speech and Gesture Interaction in an Ambient Assisted Living Lab. In *Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments* (Jeju, Republic of Korea) *(SMIAE '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 18–27.

[3] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2010. Predicting the Cost of Error Correction in Character-Based Text Entry Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1753326.1753329

[4] Muhammad Zeeshan Baig and Manolya Kavakli. 2018. Qualitative analysis of a multimodal interface system using speech/gesture. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, IEEE, Wuhan, China, 2811–2816.

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

[6] Marie-Luce Bourguet. 2006. Towards a taxonomy of error-handling strategies in recognition-based multi-modal human–computer interfaces. *Signal Processing* 86, 12 (2006), 3625–3643.

[7] Marie-Luce Bourguet and Akio Ando. 1998. Synchronization of Speech and Hand Gestures during Multimodal Human-Computer Interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems* (Los Angeles, California, USA) *(CHI '98)*. Association for Computing Machinery, New York, NY, USA, 241–242. https://doi.org/10.1145/286498.286726

[8] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola, and Ivan Poupyrev. 2004. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., USA.

[9] Joshua Brustein. 2018. Microsoft Wins $480 Million Army Battlefield Contract. https://www.bloomberg.com/news/articles/2018-11-28/microsoft-wins-480-million-army-battlefield-contract

[10] Sarah Buchanan, Bourke Floyd, Will Holderness, and Joseph J. LaViola. 2013. Towards User-Defined Multi-Touch Gestures for 3D Objects. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (St. Andrews, Scotland, United Kingdom) *(ITS '13)*. Association for Computing Machinery, New York, NY, USA, 231–240. https://doi.org/10.1145/2512349.2512825

[11] Sébastien Carbini, Lionel Delphin-Poulat, L Perron, and Jean-Emmanuel Viallet. 2006. From a wizard of Oz experiment to a real time speech and gesture multimodal interface. *Signal Processing* 86, 12 (2006), 3559–3577.

[12] Joyce Y. Chai and Shaolin Qu. 2005. A Salience Driven Approach to Robust Input Interpretation in Multimodal Conversational Systems. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Vancouver, British Columbia, Canada) *(HLT '05)*. Association for Computational Linguistics, USA, 217–224. https://doi.org/10.3115/1220575.1220603

[13] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User Elicitation on Single-Hand Microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3403–3414. https://doi.org/10.1145/2858036.2858589

[14] Aurélie Cohé and Martin Hachet. 2012. Understanding User Gestures for Manipulating 3D Objects from Touchscreen Inputs. In *Proceedings of Graphics Interface 2012* (Toronto, Ontario, Canada) *(GI '12)*. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 157–164. http://dl.acm.org/citation.cfm?id=2305276.2305303

[15] Andrea Corradini and Philip R Cohen. 2005. On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. , 97–112 pages.

[16] Andreas Dünser, Raphaël Grasset, Hartmut Seichter, and Mark Billinghurst. 2007. *Applying HCI principles to AR systems design.* University of Canterbury. Human Interface Technology Laboratory., New Zealand.

[17] Susan Goldin-Meadow, Martha Wagner Alibali, and R Breckinridge Church. 1993. Transitions in concept acquisition: using the hand to read the mind. *Psychological review* 100, 2 (1993), 279.

[18] Susumu Harada, Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. 2013. Characteristics of Elderly User Behavior on Mobile Multi-touch Devices. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 323–341.

[19] A G Hauptmann. 1989. Speech and gestures for graphic image manipulation. *ACM SIGCHI Bulletin* 20, SI (1989), 241–245.

[20] Alexander G Hauptmann and Paul McAvinney. 1993. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies* 38, 2 (1993), 231–249.

[21] Sylvia Irawati, Scott Green, Mark Billinghurst, Andreas Duenser, and Heedong Ko. 2006. An Evaluation of an Augmented Reality Multimodal Interface Using Speech and Paddle Gestures. In *Proceedings of the 16th International Conference on Advances in Artificial Reality and Tele-Existence* (Hangzhou, China) *(ICAT'06)*. Springer-Verlag, Berlin, Heidelberg, 272–283. https://doi.org/10.1007/11941354_28

[22] Robert J.K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. 2008. Reality-Based Interaction: A Framework for Post-WIMP Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/1357054.1357089

[23] 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 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (Madrid, Spain) *(ACL '98/EACL '98)*. Association for Computational Linguistics, USA, 281–288. https://doi.org/10.3115/976909.979653

[24] Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 12–19. https://doi.org/10.1145/958432.958438

[25] A A Karpov and R M Yusupov. 2018. Multimodal Interfaces of Human–Computer Interaction. *Her. Russ. Acad. Sci.* 88, 1 (Jan. 2018), 67–74.

[26] Spencer D Kelly, Asli Ozyürek, and Eric Maris. 2010. Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychol. Sci.* 21, 2 (Feb. 2010), 260–267.

[27] Sumbul Khan and Bige Tunçer. 2019. Gesture and speech elicitation for 3D CAD modeling in conceptual design. *Automation in Construction* 106 (2019), 102847.

[28] Kenrick Kin, Maneesh Agrawala, and Tony DeRose. 2009. Determining the Benefits of Direct-Touch, Bimanual, and Multifinger Input on a Multitouch Workstation. In *Proceedings of Graphics Interface 2009* (Kelowna, British Columbia, Canada) *(GI '09)*. Canadian Information Processing Society, CAN, 119–124.

[29] David B. Koons, Carlton J. Sparrell, and Kristinn Rr. Thorisson. 1998. *Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 53–64.

[30] Minkyung Lee and Mark Billinghurst. 2008. A Wizard of Oz Study for an AR Multimodal Interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/1452392.1452444

[31] Minkyung Lee, Mark Billinghurst, Woonhyuk Baek, Richard Green, and Woontack Woo. 2013. A usability study of multimodal input in an augmented reality environment. *Virtual Real.* 17, 4 (Nov. 2013), 293–305.

[32] Daniel P Loehr. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3, 1 (2012), 71–89.

[33] David Mcneill. 2005. *Gesture and Thought.* the University of Chicago Press, USA. https://doi.org/10.7208/chicago/9780226514642.001.0001

[34] Mark Micire, Munjal Desai, Amanda Courtemanche, Katherine M. Tsui, and Holly A. Yanco. 2009. Analysis of Natural Gestures for Controlling Robot Teams on Multi-touch Tabletop Surfaces. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Banff, Alberta, Canada) *(ITS '09)*. ACM, New York, NY, USA, 41–48. https://doi.org/10.1145/1731903.1731912

[35] Christophe Mignot, Claude Valot, and Noëlle Carbonell. 1993. An Experimental Study of Future "Natural" Multimodal Human-Computer Interaction. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 67–68. https://doi.org/10.1145/259964.260075

[36] Lisette Mol and Sotaro Kita. 2012. Gesture structure affects syntactic structure in speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34. CogSci, USA, 761 – 766.

[37] Meredith Ringel Morris. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces* (Cambridge, Massachusetts, USA) *(ITS '12)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2396636.2396651

[38] Meredith Ringel Morris, Jacob O. Wobbrock, and Andrew D. Wilson. 2010. Understanding Users' Preferences for Surface Gestures. In *Proceedings of Graphics Interface 2010* (Ottawa, Ontario, Canada) *(GI '10)*. Canadian Information Processing Society, CAN, 261–268.

[39] Tomer Moscovich and John F. Hughes. 2008. Indirect Mappings of Multi-Touch Input Using One and Two Hands. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1275–1284. https://doi.org/10.1145/1357054.1357254

[40] Miguel A Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets.

[41] Michael Nielsen, Moritz Störring, Thomas B. Moeslund, and Erik Granum. 2004. A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*, Antonio Camurri and Gualtiero Volpe (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 409–420.

[42] F. R. Ortega, A. Galvan, K. Tarre, A. Barreto, N. Rishe, J. Bernal, R. Balcazar, and J. Thomas. 2017. Gesture elicitation for 3D travel via multi-touch and mid-Air systems for procedurally generated pseudo-universe. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Los Angeles, CA, USA, 144–153.

[43] F. R. Ortega, K. Tarre, M. Kress, A. S. Williams, A. B. Barreto, and N. D. Rishe. 2019. Selection and Manipulation Whole-Body Gesture Elicitation Study In Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Osaka, Japan, Japan, 1723–1728.

[44] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. *Commun. ACM* 42, 11 (Nov. 1999), 74–81. https://doi.org/10.1145/319382.319398

[45] Sharon Oviatt. 2000. Taming recognition errors with a multimodal interface. *Commun. ACM* 43, 9 (2000), 45–51.

[46] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (State College, PA, USA) *(ICMI '04)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/1027933.1027957

[47] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 415–422. https://doi.org/10.1145/258549.258821

[48] Helge Petersson, David Sinkvist, Chunliang Wang, and Örjan Smedby. 2009. Web-based interactive 3D visualization as a tool for improved anatomy learning. *Anatomical sciences education* 2, 2 (2009), 61–68.

[49] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. 2014. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Munich, Germany, 73–82.

[50] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-Defined Gestures for Augmented Reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 955–960. https://doi.org/10.1145/2468356.2468527

[51] Thomas Plank, Hans-Christian Jetter, Roman Rädle, Clemens N. Klokmose, Thomas Luger, and Harald Reiterer. 2017. Is Two Enough?: ! Studying Benefits, Barriers, and Biases of Multi-Tablet Use for Collaborative Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 4548–4560. https://doi.org/10.1145/3025453.3025537

[52] Sandrine Robbe. 1998. An Empirical Study of Speech and Gesture Interaction: Toward the Definition of Ergonomic Design Guidelines. In *CHI 98 Conference Summary on Human Factors in Computing Systems* (Los Angeles, California, USA) *(CHI '98)*. Association for Computing Machinery, New York, NY, USA, 349–350. https://doi.org/10.1145/286498.286815

[53] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined motion gestures for mobile interaction.

[54] Emanuel A Schegloff. 1984. On some gestures' relation to talk.(pp. 266-296) In J. Maxwell and J. Heritage (Eds.) Structures of social action.

[55] Katherine Tarre, Adam S. Williams, Lukas Borges, Naphtali D. Rishe, Armando B. Barreto, and Francisco R. Ortega. 2018. Towards First Person Gamer Modeling and the Problem with Game Classification in User Studies. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (Tokyo, Japan) *(VRST '18)*. ACM, New York, NY, USA, Article 125, 2 pages. https://doi.org/10.1145/3281505.3281590

[56] Theophanis Tsandilas. 2018. Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *ACM Trans. Comput. Hum. Interact.* 25, 3 (June 2018), 18.

[57] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1325–1334. https://doi.org/10.1145/2702123.2702223

[58] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2016. Between-Subjects Elicitation Studies: Formalization and Tool Support. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3390–3402. https://doi.org/10.1145/2858036.2858228

[59] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob A Wobbrock. 2020. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies. In *Proceedings of ACM Int. Conf. on Designing Interactive Systems (DIS'20)*. ACM Press, Eindhoven, NA.

[60] Jacob O Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A Myers. 2005. Maximizing the guessability of symbolic input.

[61] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 1083–1092.

[62] Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller. 2011. Taxonomy of Microinteractions: Defining Microgestures Based on Ergonomic and Scenario-dependent Requirements. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part I* (Lisbon, Portugal) *(INTERACT'11)*. Springer-Verlag, Berlin, Heidelberg, 559–575. http://dl.acm.org/citation.cfm?id=2042053.2042111

[63] Ionuţ-Alexandru Zaiţi, Ştefan-Gheorghe Pentiuc, and Radu-Daniel Vatavu. 2015. On free-hand TV control: experimental results on user-elicited gestures with Leap Motion. *Pers. Ubiquit. Comput.* 19, 5 (Aug. 2015), 821–838.

Received July 2020; revised August 2020; accepted September 2020