This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3023566, IEEE Transactions on Visualization and Computer Graphics

Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation

Adam S. Williams, *Student Member, IEEE*, Jason Garcia, and Francisco Ortega, *Affiliate Member, IEEE*

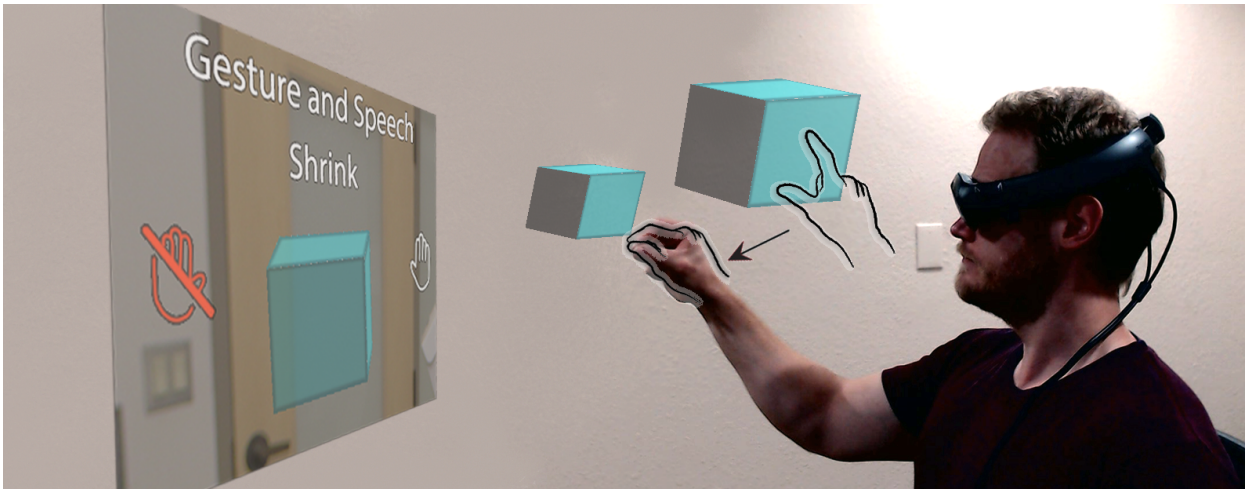

Figure 1. Example of experiment design: Left: participant view, Middle (hand outlines): gesture used, Right: Participant

Abstract—The primary objective of this research is to understand how users manipulate virtual objects in augmented reality using multimodal interaction (gesture and speech) and unimodal interaction (gesture). Through this understanding, natural-feeling interactions can be designed for this technology. These findings are derived from an elicitation study employing Wizard of Oz design aimed at developing user-defined multimodal interaction sets for building tasks in 3D environments using optical see-through augmented reality headsets. The modalities tested are gesture and speech combined, gesture only, and speech only. The study was conducted with 24 participants. The canonical referents for translation, rotation, and scale were used along with some abstract referents (create, destroy, and select). A consensus set of gestures for interactions is provided. Findings include the types of gestures performed, the timing between co-occurring gestures and speech (130 milliseconds), perceived workload by modality (using NASA TLX), and design guidelines arising from this study. Multimodal interaction, in particular gesture and speech interactions for augmented reality headsets, are essential as this technology becomes the future of interactive computing. It is possible that in the near future, augmented reality glasses will become pervasive.

1 Introduction

Understanding multimodal interaction within augmented reality (AR) head-mounted displays (HMDs) is an important step towards improving user interactions. When used as unimodal inputs gestures and speech each have their strengths [40]. Gestures can be beneficial for direct manipulation of virtual objects where speech can be beneficial for abstract tasks such as creating new objects. The combination of gesture and speech, abundant in everyday life, can provide richer information than using either of those modalities alone. The synergies and individual merits of these modalities have not yet been fully examined in AR-HMD environments. Consider the impact that the desktop computer, smartphone, and tablet have had on people's lives. Augmented reality is one of the key technologies expected to have similar impacts on people's lives. As such, understanding the best inputs and combinations of inputs for use in this emerging technology is necessary. Unlike multi-touch devices, as of now, there exists no clear standard when it comes to mid-air gestures for use in AR environments [13].

The primary objective of this research is to understand how people naturally manipulate virtual objects in AR environments using multimodal interactions (gesture and speech) and unimodal interactions (gesture). This is done by observing participants perform these interactions in an unconstrained environment. All inputs within each modality were accepted (i.e. any mid-air gesture or utterance). Given the nature of combining gesture with speech, speech alone was also examined. This addition allowed for a better analysis of how speech is formed with and without gestures. A secondary goal of this research is to assist in understanding how existing knowledge about gesture and speech interactions from psychology [27, 36, 39] hold once technology (in particular, AR) is added to the equation. Thus helping bridge the existing knowledge on human to human communication with human to computer communication.

End users represent a broad range of preferences. While most users prefer multimodal gesture and speech interactions, some users will prefer speech alone, or gesture alone [11]. With these varying individual preferences implementing gesture and speech alone as well as combined is important.

1.1 Contributions

The main contributions of this paper are:

1. A novel within-subjects multimodal and unimodal elicitation

• *The authors are with the Computer Science Department, Colorado State University, Fort Collins, Colorado. E-mail:AdamWil, J.S.Garcia, F.Ortega@colostate.edu*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

study for object manipulation tasks in optical see-through AR-HMDs (Setup seen in Figure 1).

2. Gesture only (producing a gesture set) and speech only elicitation study to highlight the individual strengths of these input modalities and a co-occurring gesture and speech elicitation study to highlight the synergies found when combining those modalities.

3. We present findings on the timing windows and syntax of co-occurring gesture and speech interactions and compare that with the syntax used in speech only interactions.

4. Design guidelines for AR interactions based on the synergies and individual strengths of gesture and speech interactions.

1.2 Multimodal Elicitation

In contrast to multimodal fusion designs, where input recognition and integration is often tested [7], we used participatory design guidelines [64] to work with the users to find which interactions they would naturally want to use. This information can be used to help improve recognizer systems' accuracy and design user-centric interactions within AR-HMD building environments.

2 Why Gestures and Speech?

Interface design must be intuitive [45]. There is a large body of work on gesture and speech in human to human communication [27, 36, 39], and human computer communication [4, 12, 34]. An interface that mirrors human to human interactions could reduce the learning time needed for technology use. With that in mind, it is important to have systems with multimodal (e.g., gesture and speech combined) as well as unimodal (e.g., gesture or speech alone) interaction capabilities. Gestures and speech together constitute language [36]. They have bidirectional influence and obligatory influence on each other, which is to say that people typically consider both at the same time [27].

Using multimodal inputs has many benefits, particularly when dealing with gestures and speech combined. Gesturing when co-occurring with speech has been shown to help lower the cognitive load of a task [16], there are hints at sped up task completion time, and even lower error rates [34]. Each information stream (gesture, speech) contains non-redundant information [15] which can facilitate the disambiguation of the inputs from the other channel [25, 30, 49].

Given the option of using gestures, speech, or both combined participants used both 60% to 70% of the time [12, 19]. This can be exploited to help improve recognition accuracy [14]. Users feel that interactions are more natural when they have multiple input modalities and can choose the one that best suits them [3, 26]. The ability to have true multimodality could further improve their interactions.

Current AR-HMDs (i.e. Magic Leap One and Microsoft Hololens) are built with gesture sets that are limited and likely designed for recognition accuracy, not ease of use. For example, Magic Leap's "C" gesture is fairly easy to detect (being a static symbolic gesture) but may not be the most natural. Additional examples can be found in the other default gestures for the Magic Leap One and the Microsoft HoloLens 1. Occasionally gesture sets are derived from users; however, these are often expert users [65]. People typically prefer user-defined gesture sets to expert-designed sets [64]. There is also evidence that elicited gestures are up to 24% more memorable [44].

The gap between traditional input devices and combined gesture and speech inputs is being minimized by advances in technology, soon gesture and speech inputs will be more efficient than traditional input devices [3].

Switching to these modalities is no trivial task. When using AR-HMDs, issues include gestures for ego-centric cameras such as the head mounted cameras on most HMDs, self-occlusion, device field of view (FOV), natural feeling interactions, common speech mappings, and timings of co-occurring gestures and speech when in virtual environments. This work tackles some of those issues and provides information on the individual and joint strengths of these modalities, a consensus gesture set, co-occurring gesture and speech timing information, and design guidelines to use when developing building applications for optical see-through augmented reality head-mounted displays.

3 Previous Work

Gesture elicitation is a study design that can help us map gestures to actions for emerging technologies. The elicited inputs have the goal of being highly discoverable to novice users of systems [64]. Elicitation studies also allow us to better understand user behavior. Elicitation studies have found that people use larger motions for larger objects when attempting the same action [52, 57], and that there is a preference for upper-body gestures even when a whole-body system is available [47]. Most commonly these studies have been conducted using Wobbrock et al.'s methods [63, 64], later refined by Vatavu and Wobbrock [59, 60] (variations exist [61]). This study used gesture elicitation, as well as mutlimodal gesture and speech elicitation, which is less common [28, 40].

3.1 Gesture Elicitation

These methods normally include the use of a Wizard of Oz (WoZ) experiment design. WoZ experiment design is a way to remove the gulf of execution between the participant and the system [64]. In a WoZ elicitation experiment, a participant is shown a command to execute such as *move left*. This command is called a referent. Then the participant provides some sort of input proposal for that referent and behind the curtain, so to speak, an experimenter triggers the recognition of that input. In the experiment presented here that would look like a participant proposing a gesture (in the gesture modality) to move a virtual object left, then the experimenter, upon seeing this proposal, triggering the movement of the object. In this way, inputs can be designed for emerging technologies without perfect recognizers existing. After all the input proposals are collected they are binned into equivalence classes and measures of consensus between participants are used to generate input set proposals. This process is elaborated on later.

Many follow-up studies have created gesture sets using gesture elicitation [6, 9]. The popularity of gesture elicitation can be seen in the variety of the studies that use it, from multi-touch surfaces [6, 37], and mobile devices [54], to internet of things home sets ups [66]. Efforts to enhance further elicitation studies have led researchers to devise alternatives that extend beyond surface-computing devices, such as using multi-touch and mid-air devices in tandem [52, 62] and using multi-touch devices to control physical objects through virtual representations of said entities [17]. Imposing constraints on the users' motion has also led to new elicitation studies primarily concerned with defining and investigating gesture sets suitable for both impaired and non-impaired users [1, 55].

3.2 Gesture and Speech Studies

Gesture and speech input modalities have been studied for some time. Many studies have looked at ways of combining them as input channels using multimodal fusion modals [4, 7, 24, 50]. The goal of those studies was to implement recognition systems. Studies have also looked at the timing windows of co-occurring gestures and speech [33]. There is work on the usability of limited gesture sets [8] and constrained speech dictionaries [53]. Those types of works are aimed at understanding some combination of the feasibility of gesture and speech inputs, the adaptability of people to constrained inputs, and the implementation of fusion models for gesture and speech recognition. Those works typically start with defined acceptable inputs, maybe "open palm swiping" in the case of gestures [8], then test usage from there.

The work presented here is very different in that there are no constraints imposed on input proposals. Participants are free to generate any proposal that they feel is best suited to the referent displayed. There have been previous studies on gesture and speech interactions. Table 1 shows a list of studies that use WoZ methods to observe or elicit gestures and speech interactions. Most of those studies did not have the goal of generating a consensus set of inputs. While a few of them did observe mid-air gestures [2, 7, 20, 28, 33, 40], some only looked at a subset of gesturing such as pointing gestures [5, 53], paddling gestures [23], or 2 dimensional (2D) gestures [38, 53]. The work presented

Table 1. Previous gesture with speech elicitation studies

Authors	Display used	Consensus set made	Paired elicitation	Use case	Gestures accepted	Independent testing of modalities
Hauptmann et al. [20]	2d Screen	No	No	Graphic manipulation	Mid-air	Yes
Mignot et al. [38]	2d Screen	No	No	Control a process	Touch	No
Bourguet [53]	2d Screen	No	No	Explanations of process	Pointing	No
Carbini et al. [7]	2d Screen	No	Yes	Tell a story	Mid-air	No
Lee et al. [33]	2d Screen / handheld AR	No	No	Object manipulations	Mid-air	No
Morris [40]	2d Screen	Yes	Yes	Web browsing	Mid-air	No
Anastasiou et al. [2]	Room	No	No	Accessibility	Mid-air	No
Robbe [53]	2d Screen	No	No	Constrained speech dictionary	Touch miming and pointing	No
Khan et al. [28]	2d Screen	Yes	No	Computer aided design	Mid-air	Gesture / gesture or speech
Irawait et al. [23]	Optical see-through AR-HMD	No	No	Object manipulations	Open hand gestures	Gesture / gesture with speech
The study presented here	Optical see-through AR-HMD	Yes	No	Object manipulations	Mid-air	Yes

Legend: AR: Augmented Reality, HMD: Head mounted display, Miming gestures: charade like gestures

here examines any gesture and / or utterance that a participant feels is appropriate for a given referent.

The study that is most similar to this is a gesture and speech elicitation study done for developing commands for a television-based web browser [40]. Participants were placed in paired elicitation sessions where the dyads of participants made proposals together. The referents were read out loud to the participant. For the referent *move left* the experimenter would read "move left". Participant dyads were given the choice of using either gesture, speech, or both; however, the modalities were not tested individually. Commands for web browsing on a television (i.e. "refresh page") are decidedly different from the commands needed to manipulate objects in optical see-through AR environments.

A second similar study did gesture and speech elicitation for computer-aided design (CAD) programs to be used with 2D screens [28]. This experiment tested gesture alone, then gesture with speech. They provide a consensus set of utterances and gestures. This study chose to show the referents' action in the form of an animation as opposed to as text. For the referent *move left* the participant would see the virtual object moving left. This study is domain-specific to CAD program usage. Previous work has found that prompting users to gesture with 2D screens compared to 3 dimensional (3D) objects can impact the production of gestures [10]. Additionally, Khan et al. informed users that they were describing referents to another person though use of a video system. The notion of describing a referent to a person compared to executing a referent in a system is an important distinction. This work also extends the work of Khan et al. by providing the timing information of co-occurring gesture and speech interactions.

All of the studies shown in Table 1 have furthered the field of gesture and speech interactions. Still, those studies are different from the work presented here in some major ways. Including the pairing of participants, domains of application, and how the referents are presented. Most of those studies only tested interactions in a single pass where participants proposed speech alone, gesture alone, or both together. Whereas this paper tests each modality independently. The last row of Table 1 shows the methods used in this study, to be compared with the other works. This study will help to further gesture and speech elicitation methods, AR-HMD interactions, object manipulations in 3D space, and finding differences between when speech alone is used and when co-occurring gestures and speech are used.

While the research presented here is not on gesture recognition or multimodal input fusion, elicitation can provide important findings for future recognizers (including findings from this research). Recognition of gestures has been attempted in many ways; however, it has not often been done with AR-HMD's and ego-centric cameras.

3.3 Elicitation Criticisms

Elicitation methods have received criticism in two major areas. First, it was suggested that common consensus metrics were too permissive because they do not account for the base chance of randomly selecting a proposal for a given referent [58]. Tsandalis proposed using Fleiss' kappa and a chance agreement term in addition to those metrics to address this [58]. We have analyzed our data using those statistics to alleviate this concern. Second, there is a concern that given the exposure to existing devices or gestures, elicitation may be biased (i.e., legacy bias). This has been examined, and various ways to incorporate it [31, 47] or reduce it [41, 55] have been introduced. However, other than priming [22], no reduction methodology has shown promise, except for physical constraints [8], but constraints are infeasible in some cases. Some work has shown that legacy bias can be beneficial in finding gestures for abstract tasks [51].

4 Methods

This work performed an elicitation study using the WoZ methods to find natural feeling gesture, speech, and gesture with speech interactions for the manipulation of rendered 3D objects in optical see-through AR environments. The input modalities used were Gestures (G), Speech (S), and Gesture with Speech (GS). Each modality was tested independently in a within-subjects experiment design. Our methodology is derived from our previous work and the literature already described. These include agreement rate ($\mathcal{AR}$)[1], co-agreement rate ($\mathcal{CR}$), and the V_{rd} significance test [59,60,64]. When reporting overall agreement rates for gesture proposals, we also make use of Fliess's Kappa coefficient (κ_F) and the chance agreement term (p_e) as described by Tsandilas [58].

Both speech and gesture proposals were annotated based on the video data from the exo-centric and ego-centric cameras. Proposals then were binned into equivalence classes by the experimenter. Gestures were binned based on the direction of movement, and hand pose. Hand poses were "grasping" where all fingers were closed, "pinching" where just the thumb and index or thumb index and middle fingers were touching, "open" where all fingers were extended, and "index finger" where only the index finger was extended. Previous work showed that users care less about the count of fingers used than the hand pose used [62]. Movements were based on the axis of movement. For example, translations right and left were both considered movements on the y-axis. If a gesture could not be binned in this manner it was given its own class(i.e. tracing a square). For speech calculations,

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

words were binned only if they were nearly identical. Saying "move forwards" and "move forward" were considered the same where "move towards" would be different.

The original metric for consensus is the agreement index which involves the proportion of participants proposing equivalent gestures [63]. This metric was changed to $\mathcal{AR}$ which addresses some of the issues with the original formula, adjusting the output values to between 0 and 1 [59]. $\mathcal{CR}$ is defined as a measure of shared agreement between two referents. It is calculated as the count of pairs of participants that are in agreement for two referents over the total possible pairs of participants [59]. For speech alone the consensus-distinct ratio (CDR) was used. The CDR is the percent of equivalent proposals given by more than two participants for a given referent [40].

4.1 Participants

The study consisted of 24 volunteers (4 female, 20 male). Participants were recruited using emails and through word of mouth. Ages ranged from 18 - 43 years (Mean = 23.32, SD = 5.23). All participants reported heavy computer usage but limited video game usage. Two participants were left-handed. Eleven participants reported less than 30 minutes of Microsoft Hololens 1 usage before this experiment. Seven participants learned English as a second language and reported fluency in English.

4.2 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 was recorded on the Magic Leap One. In addition, we used a GoPro hero 7 black (to record an ego-centric view of the interactions) and a 4k camera (to record an exo-centric view of the interactions). Each referent was shown 50 centimeters in front of the user. Users were given an on-screen aid to tell if their hand/hands were inside of the FOV of the device. This aid showed one hand on each side of the screen in red unless a hand was seen. If a hand was sensed the corresponding aid (left to left, right to right) would turn white, as shown in Figure 2.

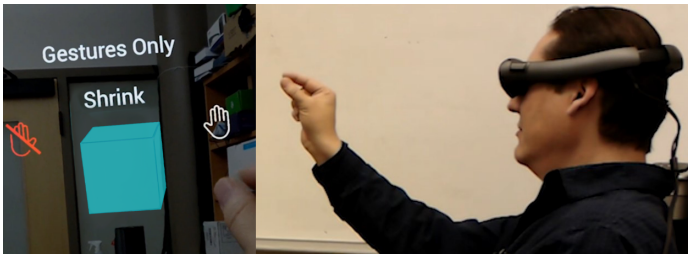

Figure 2. Left:participant view in experiment, Right: participant

4.3 Referents

Table 2. Referents used by category

Translation	Rotation	Abstract	Scale
Move (Left / Right)	Roll (C / CC)	Create	Enlarge
Move (Up / Down)	Yaw (Left / Right)	Destroy	Shrink
Move (Towards / Away) from self	Pitch (Up / Down)	Select	

Legend: C: Clockwise; CC: Counter Clockwise

Referents (i.e. actions) for canonical manipulations [32] including selection, scaling, translation (on x,y, and z axes), and rotation (about x,y, and z axes) were used. In addition, application-specific manipulations [32] which included create and delete were used. All the referents are listed in Table 2. The goal of this study is to create an interaction set for object manipulations in any sort of virtual environment that uses building tasks (e.g., Lego-like applications). Specifically, when AR-HMDs, egocentric viewing, and multimodal inputs are used. Object selection was tested independently in the *select* referent. For the other referents, participants were told that they could assume the object was already selected. Referents were displayed as text. This decision was informed by previous work [28, 40] and the results of pilot studies which are discussed further in the *Results* section.

4.4 Procedure

At the start of each session, participants completed an informed consent and demographics questionnaire. The questionnaire included questions about prior device usage, game usage, and handedness. Participants were then shown a 2-minute video with the instructions for the experiment. They were informed that they would be asked to complete a series of object manipulations using different modalities of input and within each modality (G, S, or GS) they could use whatever input they wished. For example, if the modality was gesture than any gesture proposed was accepted. Participants were then given a practice trial for each of the modalities. During this time, they were invited to ask questions, adjust the device, and play with the device's gesture sensing range using the on-screen hand detection aid, see the left side of Figure 2. Note that the gesture sensing was to add realism to the experiment but this experiment was a WoZ elicitation experiment.

Participants were presented the interaction modalities based on a Latin square division of blocks. For example, participants may have seen speech first, then after completing all the referents for speech, see the next modality (G or GS in this example). Referents were shown in random order. The object to be manipulated was a cube rendered approximately 50cm in front of the user. A cube was chosen to allow visual ques of rotations which would be more difficult to see with either a sphere or cylinder. A cube represents a basic object that most users have interacted with in the real world. Using a cube limited some of the object specific grips that could appear in interactions with complex shapes (hand pose matching uneven object surface).

The referent was shown as a text banner and above that, the interaction modality requested was shown. On either side of the cube was a hand that was either red with a line crossing it or white (left side of Figure 2). The hands indicated whether or not a participant's hand was in the camera's field of view. An example of a referent and corresponding gesture proposal is shown in Figure 2. After a proposal was made by the participant the virtual object would execute that referent and the next referent would be loaded. In the G and S blocks this execution occurred when any proposal was given. To ensure constancy of proposal modality in the GS block both an utterance and a gesture had to be proposed before the referent was executed. After each interaction modality, the NASA TLX [18] survey was administered.

5 Results

The agreement rate ($\mathcal{AR}$), co-agreement rate ($\mathcal{CR}$), and (V_{rd}) statistic were used to quantify consensus among participants. Fliess's Kappa coefficient (κ_F) and the associated chance agreement term (p_e) [58] were used when reporting the overall agreement rates for the gesture proposals. Where applicable, the appropriate statistics were computed using the AGATe 2.0 tool (AGreement Analysis Toolkit) [2]. For the speech proposals, the consensus-distinct ratio (CDR) was used [40].

The agreement rate $\mathcal{AR}$ is defined as the number of pairs of participants in agreement with each other divided by the total number of pairs of participants that could be in agreement. Shown formally for a single referent r in Equation 1, where P is the set of all proposals for referent r, and P_i are the subsets of equivalent proposals from P.

$$\mathcal{AR}_r = \frac{\sum_{P_i \subseteq P} \frac{1}{2}|P_i|(|P_i|-1)}{\frac{1}{2}|P|(|P|-1)} \quad (1)$$

[2] Available at http://depts.washington.edu/acelab/proj/dollar/agate.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3023566, IEEE Transactions on Visualization and Computer Graphics

5.1 Pilot Studies

Two versions of this elicitation experiment were run on pilot groups consisting of 6 people each. In one, we displayed the referents as text (Figure 2), in the other we showed the action of the referent then asked for proposals. As an example, if the referent was *move left*, in the first set up the screen read "move left" and participants were asked to propose a command to execute that referent (similar to [40, 64]). Upon generation of that proposal, the virtual object would move. In the second design, the virtual object would move then participants were asked to generate an appropriate command proposal (similar to [28]).

During the speech block of the pilot study where referents were displayed as text participants would commonly repeat the referent displayed. If the referent was *move left* the utterance was also "move left". This is not entirely unreasonable. In the pilot study without text, for simple translations, the most frequent utterances were "move" and the direction such as "left". This repeating of referents, either the entire referent or a sub-portion of it can also be seen in the results of Morris [40]. An example from that study is that when given the referent *open new tab* the top utterances were "new tab" and "open new tab".

In the version with referents shown as movement, people would nearly always propose a gesture that was as close as to a one to one manipulation with the object's motion as was possible. For rotations, people would twist their wrist into uncomfortable positions to try and match the object's motion. For the abstract referents, people's gestures would mirror whatever animation was shown. If the virtual object was materializing from right to left, their hand moved from right to left. More troublingly, none of the participants understood what was being asked of them when the referent was *create* and the virtual object appeared with no animation. The effect referent animations biasing gesture production can be seen in [28]. Examples from that study include the proposed gestures for the *orbit* and *pan* referents which have participants' top choice gestures mirroring the visual motion of those referents.

Due to the evidence of priming gestures found when showing the referent as an animation, we have chosen to display referents as text. This set up can be seen in Figure 2. There is no perfect solution for how this experiment should be run. The text banners had less priming on the gesture alone and the gesture and speech conditions. In the case of speech alone, some speech was primed to repeat the referent as displayed, also seen in [40]. This was not always the case. For some referents, such as the rotational referents, the utterances "tilt", "rotate", and "spin'' occurred with high frequency. These utterances were also found in the pilot study where users were shown the animation of the referent with no text. We believe that while the individual utterances found in the speech block should be observed skeptically, the overall results still yield insights into what utterance people will gravitate towards using in augmented reality manipulation tasks, as Morris's work with similar biasing yielded insights into appropriate speech commands for web-browsing on large screen displays [40]. Additionally, because speech was examined alone, differences in what utterances occur alone versus what utterances occur when accompanied by gestures can be examined.

5.2 Gesture Only Block

The overall agreement rate observed for the Gesture block was .353 with $\kappa_F = .317$. The low chance agreement term ($p_e = .058$) used in Fleiss's Kappa coefficient indicates an agreement beyond chance [58], allowing us to consider rates above 0.3 as high levels of consensus between participants given our N of 24 based on the simulations of varying agreement distributions found in [59]. The agreement rates for each referent are given in Figure 3 and shown as numbers in Table 6 and Table 5.

The effect of referent type on agreement rates was observed to be significant ($V_{rd(16,N=408)} = 856.872, p < .001$). The highest single agreement rate belonged to the referent *Select* ($\mathcal{AR}_{Select} = .837$), which may be due to the legacy bias from the smart phone (e.g., iPhone). The more abstract referents, *Create* and *Delete*, exhibited extremely low agreement rates ($\mathcal{AR}_{Create} = .083$, $\mathcal{AR}_{Delete} = .08$).

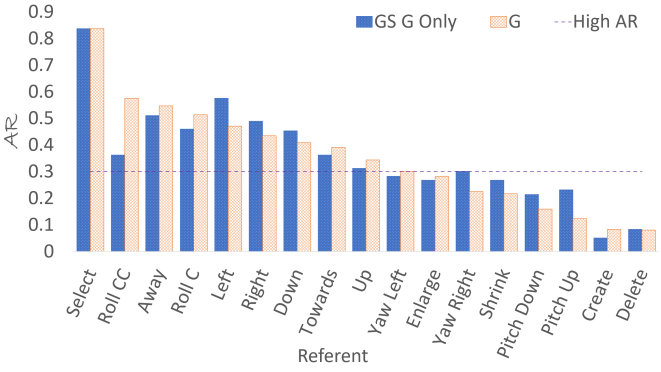

Figure 3. Agreement rates for gestures in the gestures block (G) and the gesture with speech block (GS); C: Clockwise; CC: Counter Clockwise

The referents involving a physical translation (*up, down, left, right, away*, and *towards*) had high gesture agreement among participants (average $\mathcal{AR} = .433$). Among these translational referents, the direction of motion displayed a significant effect on agreement rates ($V_{rd(5,N=144)} = 41.446, p < .001$), with *away* achieving the highest individual agreement ($\mathcal{AR}_{Away} = .547$). While no significant difference in agreement was found between *right* and *left* ($V_{rd(1,N=48)} = 2.174, p < 1$), a significant disparity was observed for referents *towards* and *away* ($V_{rd(1,N=48)} = 18.677, p < .001$).

For the three pairs of translational referents, *Right* and *Left* had the highest co-agreement rate ($\mathcal{CR}_{Right,Left} = .37$), indicating that 37% of all participant pairs agreed on both referents.

While the average of the rotational referents (average $\mathcal{AR} = .316$) was comparable to the translational group, this was primarily due to the out sized contribution of *roll clockwise* and *roll counter clockwise*. Presumably, this high agreement for *roll* (average $\mathcal{AR}$=.545) can be attributed to the implied clock metaphor with participants pantomiming the rotation of clock hands. Among the rotational referents, the impact of referent type on agreement is considerable ($V_{rd(5,N=144)} = 271.232, p < .001$), reflecting the great disparity between *roll*'s elevated agreement and the relatively low consensus observed for *pitch* ($\mathcal{AR}_{PitchUp} = .123$, $\mathcal{AR}_{PitchDown} = .159$). Moreover, for the three pairs of rotational referents, 39% of all pairs of participants agreed on both *Roll Clockwise* and *Roll Counter Clockwise* ($\mathcal{CR}_{CW,CCW} = .391$).

It should be noted that although *Shrink* and *Enlarge* exhibited comparable agreement rates ($\mathcal{AR}_{Enlarge} = .283$, $\mathcal{AR}_{Shrink} = .217$), there was little agreement among pairs of participants for both referents ($\mathcal{CR}_{Enlarge,Shrink} = .123$).

5.3 Speech Only Block

Table 3. Consensus-distinct ratio for the speech and gesture with speech blocks by referent type

Category of referent	Gesture and Speech	Speech
Abstract	39.52%	24.52%
Rotation	44.72%	39.76%
Scale	32.50%	39.29%
Translation	53.89%	61.11%

Displaying the referent in elicitation studies [46] and reading the referent out loud in gesture and speech elicitation studies [40] both have precedence. As previously noted, these practices can prime the utterances proposed. Often the referent as displayed was repeated, however, this was not always the case. When it was, the referents were simple such as "move left". The repetition could be in part due to priming, though it could also be that there are few aliases for the phrase "move left".

The average CDR for each category of referent (Table 2) is shown in Table 3. The translations hold the highest CDR. This can be interpreted

as the translation referents having the least disagreement on the appropriate utterance proposal. Translations were nearly always the direction of movement alone (i.e. "left", "up") or a <action> <direction> pair (i.e. "move up"). The scale and rotational referents had more disagreement shown by the lower CDRs at 39.29% and 39.76% respectively. The lower CDR for scaling referents was due to a high number of aliases for each proposal, in the case of *expand* they included "grow", "zoom", and "expand". For rotations the phrases "rotate", "spin" and "tilt" paired with a direction such as "up" were proposed. "Spin" and "rotate" were commonly used for *Yaw*, "tilt" for *pitch*, and "roll" for *Roll*. "Select" was proposed by each participant for *Select*, however, there was disagreement on how to indicate the virtual object. Participants commonly said "select cube" but some said "object", or "that". The referent category with the lowest CDR was abstract referents at 24.52%. These being *create* and *destroy*. This is interpreted as meaning for the abstract tasks there was high disagreement between proposals. Commonly proposals used the word "create" or 'destroy" but disagreed on the object identifier, as seen with *select*.

We believe that aliasing commands would be beneficial when dealing with unimodal speech, as do [40, 64]. While our participants were told that they could use any utterance that they wanted, they primarily stuck to <action> <direction> or <action> <object> <direction> phrase structure. The rates for the syntax are found in Table 4. A chi-square test of independence showed that there was a significant association between block and syntax choice $X^2(2, N = 408) = 71.28, p < 0.01$. For most commands, the direction and type of manipulation were proposed (e.g., "move left", "roll right"). For commands with lower CDR we recommend aliasing some of the manipulation terms. Specifically, "spin", and "roll" were used interchangeably. For decreasing object size the combination of "smaller", "small", and "shrink" would cover 75% of proposals.

Table 4. Usage of syntax format by block

	Other	<action> <direction>	<action> <object> <direction>
GS	24.31%	62.75%	16.91%
S	11.52%	86.27%	2.21%

Legend: S: Speech block; GS: Gesture and Speech block; other: single or many word command

5.4 Multimodal Block: Gesture and Speech Combined

This section provides three analyses of the co-occurring gesture and speech block (i.e. multimodal interactions). First, the gesture portion of this block was isolated for comparison with the gesture only block (subsubsection 5.4.1). Second, the speech portion of this block was isolated for comparison with the speech alone block (subsubsection 5.4.2). Third, the gestures and speech from this block were analyzed. This breaking apart of the analyses allows for a more thorough examination of the data and better comparisons with the other modalities (previously described in subsection 5.2 and subsection 5.3).

5.4.1 Gesture in GS Block

This is the analysis of the gesture proposals alone from the GS block. The overall agreement score observed for the gestures in the GS block was .357 with $\kappa_F = .318$. The chance agreement term in Fleiss's Kappa coefficient ($p_e = .057$) indicated an agreement beyond chance [58], allowing us to consider agreement scores above 0.3 to be meaningful [59]. The agreement scores for each referent of the GS block are displayed in Figure 3.

The influence of the type of referent on the agreement rates was, again, measured to be statistically significant ($V_{rd(1,N=48)} = 770.497, p < .001$). As in the Gesture block, *Select* had the highest individual agreement rate ($\mathcal{AR}_{Select} = .837$), while again *create* and *destroy* ($\mathcal{AR}_{Create} = .051$ and $\mathcal{AR}_{Delete} = .083$) could, at best, be described as negligible agreement.

The translational referents maintained a high gesture consensus (average $\mathcal{AR}$= .451) over the GS block and the agreement rates were, again, significantly influenced by direction ($V_{rd(5,N=144)} = 87.488, p < .001$). While the referent *Left* had the highest single agreement rate ($\mathcal{AR}_{Left} = .576$), the referent *away* still retained a high consensus with $\mathcal{AR}_{Away} = .511$. The dichotomous pair (*Up,Down*) showed the largest variation in agreement with $V_{rd(1,N=48)} = 32.362, p < .001$. Not surprisingly, the pair (*Right,Left*), with a high co-agreement rate of $\mathcal{CR}_{Right,Left} = .402$, only displayed a slightly significant difference in agreement rate ($V_{rd(1,N=48)} = 8, p < .01$).

Overall agreement for the rotational referents (average $\mathcal{AR} = .309$) was lower than the translational group and the impact of referent type on consensus, while still present, was decidedly diminished ($V_{rd(5,N=48)} = 77.996, p < .001$) as compared to the gesture block. Ostensibly this can be attributed to the decreased difference between *roll*'s high agreement (average $\mathcal{AR} = .411$) and *Pitch*'s relatively low agreement (average $\mathcal{AR} = .223$).

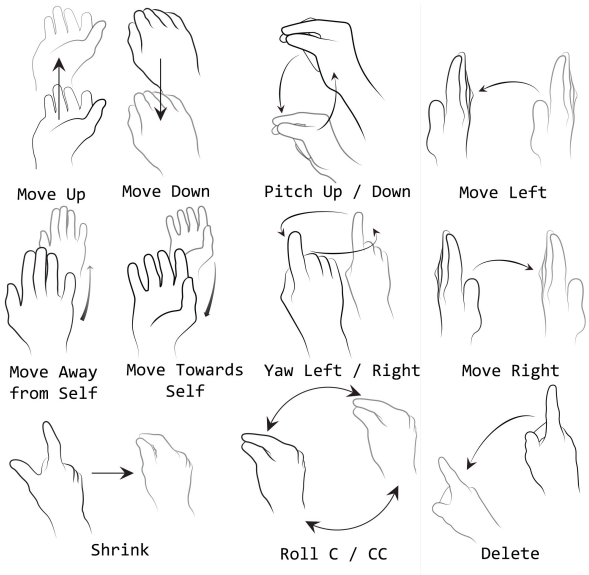

Figure 4. Proposed gesture set; C: Clockwise; CC: Counter Clockwise; Bi-directional gestures indicated with double arrows

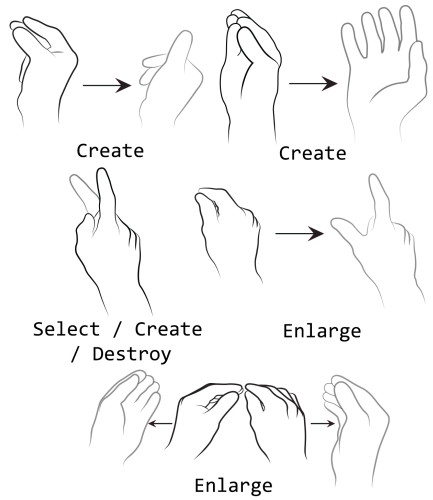

Figure 5. Gestures with ties

Table 5. Tied Gestures

Input	Referent	Gesture	$\mathcal{AR}$
GS	Create	Bloom	0.05
GS	Create	Legacy tap	0.05
GS	Delete	Legacy tap	0.08
G	Enlarge	Two Hand Grow	0.28
G	Pitch Up	Push up	0.12
GS	Pitch Up	Circle forward grabbing	0.23

Legend: C: Clockwise; CC: Counter Clockwise, G: Gesture block; GS: Gesture and Speech block

Table 6. Winning Gestures

Referent	Gesture	$\mathcal{AR}$ G	$\mathcal{AR}$ G in GS
Away	Push away	0.55	0.51
Delete	Swipe R to L	0.08	0.08
Down	Swipe down	0.41	0.45
Enlarge	Legacy zoom in	0.28	0.27
Left	Swipe left	0.47	0.58
Pitch Down	Circle forward grabbing	0.16	0.21
Right	Swipe right	0.44	0.49
Roll C	Circle C grabbing	0.51	0.46
Roll CC	Circle CC grabbing	0.58	0.36
Select	Legacy tap	0.84	0.84
Shrink	Legacy zoom out	0.22	0.27
Towards	Pull towards	0.39	0.36
Up	Push up	0.34	0.31
Yaw Left	Circle pointing up	0.3	0.28
Yaw Right	Circle pointing up	0.23	0.3

Legend: C: Clockwise; CC: Counter Clockwise; G: Gesture block; GS: Gesture and Speech block

5.4.2 Speech in GS Block

The speech alone in the GS block achieved higher CDR for most categories of referent indicating more agreement in proposals for a given referent (Table 3). This was due to less disagreement on the direction and object identifiers. Due to the pairing of gestures with speech participants would indicate the direction of movement with their finger (a finger tracing a circle in the case of yaw, see Figure 4). When using this style of command, participants would point to the object first then initiate their command. Translations CDR dropped to 53.89%. In the speech alone block participants would most commonly (86.27%) use the "move" and direction phrases together (Table 4). When gestures were also allowed participants would default to only using the direction phrase and a gesture or less commonly "move" alone and a gesture indicating the action ("Other" column in Table 4). Seen as proposing "right" and a pointing gesture. During interviews held after the experiment, most participants indicated wanting to do translations via gesture manipulations only. The same pattern is seen in the scaling referents. Participants had more disagreement with the speech to use in this condition.

What is important to note here is that the speech commands for abstract referents had less disagreement in the gesture and speech block. This indicates that while gesture alone is well suited to translations gesture with speech is more suited for abstract commands.

5.4.3 Speech with Gestures in GS Block

With the vastly larger proposal space offered when giving an utterance with a gesture the $\mathcal{AR}$ metric breaks down. This is to say that while mid-air gestures are somewhat limited in the number of proposals available, speech is far more nuanced. The combined pairings of gesture with speech are too varied for the use of $\mathcal{AR}$ without artificially binning words into equivalence classes. When observing the pairing of gesture and speech as a whole we find that 10.42% of the participants using the <action> <direction> pattern in speech used a <action> <gesture> proposal in gesture with speech. For translation referents, this looks like a participant saying "move" and swiping with their finger in a direction. With rotations, participants would say "rotate" or "spin" and tracing a circle with their finger (Figure 4).

5.4.4 Timing of co-occurring gestures and speech

The times between when a gesture was initiated and an utterance was initiated in milliseconds were (M = 151.31, SD = 120.24, Median = 130). These were measured by the time of any hand starting to move to the first sound emitted, or utterance to gesture if the utterance occurred first. This data took a non-normal distribution (Shaprio-Wilks P = 2.2e-16). We speculate that this is because on several occasions participants had to stop to think about which rotation they were performing, heavily skewing the time and causing many outliers. Based on a Wilcoxon Signed rank test ($P < 0.01$) we can assume that the true median for the data is above zero. This means that gestures are nearly always started before speech.

This result is similar to previous results [5, 33]. The results found in this study are primarily manipulative gestures whereas the results in previous work were experimenter defined deictic gestures (i.e. pointing gestures) [5] and spontaneous gestures that were primarily deictic [33]. This shows that the commonly found timing window for co-occurring gestures and speech exists for both deictic and manipulative gestures. This result also shows that gesture and speech interactions in AR-HMDS have similar timings [35] and patterns of occurrence [56] as ones outside of them.

5.5 NASA Task Load Index

The NASA TLX is a survey that is used to measure a participant's perceived workload for a given task [18]. The mean scores for the NASA TLX overall workload for the three blocks are shown in Table 7. An ANOVA showed that there is evidence of a difference between the means of the three groups (df=2,69, p = .053). We take this to mean that producing both gestures and speech combined had a higher perceived workload than producing either individually. This follows the logical intuition that producing two inputs is harder than producing one. We speculate that given an interaction set, thus not needing to create proposals, there would be lower perceived workload with multimodal inputs. As is seen in other multimodal studies [19, 21]. Admittedly a p-value of 0.053 is not equal to 0.05. That said with previous findings suggesting the same conclusion we speculate that given a larger N a difference in the overall workload would have been found.

Table 7. Average NASA TLX scores by block

	Gesture	Speech	Gesture and Speech
Mean	39.3	33.5	43.5
SD	13.4	15.6	13.3

5.6 Trial Times

The times for each trial as measured by when a referent was presented and the participant started a gesture or utterance in milliseconds are shown in Table 8. Linear contrasts showed that there is a significant difference between both gestures and speech versus gestures with speech (both $P < 0.01$, df = 1216). There was no significant difference between gesture alone and speech alone trial times (P = 0.91, df = 1216). Which follows what is expected; producing gestures and speech took longer on average than just producing gestures or speech alone. As this was measured from when either a gesture or utterance was initiated this implies that the gestures and speech block took more planning before a response.

Table 8. Average trial times by block in ms

	Gesture	Speech	Gesture and Speech
Mean	282	287	323
SD	158	158	186

5.7 Consensus Set

Most referents had a single most common gesture, seen in Figure 4. Some referents had ties shown in Figure 5. The ties for *create* predominately occurred in the GS block. All of the manipulative gestures were symmetric and bi-directional. Meaning that *roll clockwise* would be tracing a clockwise circle and *roll counterclockwise* would be tracing a counter-clockwise circle in the same manner. In the G block people swiped down and to the left for *delete* as seen in Figure 4. When speech was allowed some people switched to the taping gesture (*Select* / *Create* / *Delete* in Figure 4) and using a word for the action. A tie was found between the *enlarge* proposals where both the single hand legacy zoom in gesture and a two-handed expansion gesture were produced (Figure 5). The expansion gesture is the only two-handed gesture that occurred with enough frequency to be shown. There were a number of two-handed gestures proposed for translation that were symmetric bi-manual versions of the single-handed gesture (two hands pushing forward).

6 Discussion

In contrast to the findings of Khan et al [28]., this study found that most gesture proposals were one-handed. There were differences in the gestures produced for scaling which were predominately bi-manual hand expansions in Khan et al. and a mix of bi-manual expansions and the legacy touchscreen zoom in zoom out gesture in this work(Figure 5). The translation gestures found in this study were nearly always direct manipulation gestures. Khan et al. found bi-manual direct manipulations and bi-manual path tracing gestures for translations. Rotations were comparable between the two studies. For rotational referents the "hold and rotate" gesture found by Khan et al. was similar to the pinching roll here (Figure 4). Speech found by Khan et al. was similar to the speech found in our study for the translations where "move" was the most common choice in both studies. It is difficult to compare results for the other referents as either the axis of movement is not listed or the referents do not match.

The differences found in gestures produced between these studies could stem from the participant believing they were interacting with a human versus a system. Another cause could be the way the referents were presented to participants. Interactions with a 2D screen may be formed differently than those in 3D space [10].

When comparing to the augmented reality gesture elicitation study done by Piumsomboon et al. the translation gestures for both studies were often open handed [51]. Rotations were varied from previous work. Most rotations found here involved a pinch or index finger extended with movement following the path of a circle. Piumsomboon et al., encountered lose griped gesturing where a participant would grab the virtual object and rotate their wrist while holding it. The scaling gestures proposed in this study were commonly single handed (Figure 4) where the proposals found in Piumsomboon et al. were more often bi-manual [51]. The exception being the bi-manual "enlarge" gesture found here (Figure 5) which mirrored the uniform scale on the X-axis proposal [51]. Across both studies, most of the gestures found were reversible [51]. This is shown in the rotation and translation gestures in Figure 4.

Scaling was comparable across these two studies presumably due to participants' legacy bias from interactions with multi-touch devices(e.g. cellphones). When differences were found it could be due to the difference in the presentation of the referents. Piumsomboon et al. showed referents as animations of the intended action where this work showed referents as text.

6.1 Individual Strengths

During the practice block, participants were encouraged to move both hands in front of the device sensors to see the range of the device's hand recognition, then instructed to use one or both hands as they deemed appropriate. Even so, participants tended to use one-handed gestures (Figure 4, Figure 5). This mirrors what was found on multi-touch surfaces in [29,42,43] and mid-air full-body studies [46]. People tend prefer simple interactions over more complex ones [42]. We believe that the high number on one-handed interactions found in this study was due to the referents low level of complexity and that preference for simple interactions when possible.

Translation gestures shared high agreement rates for both the gesture and the gesture and speech blocks. Most often, participants reached forward to where the object was rendered and preformed a direct manipulation (Figure 4). For example, they reached out and pushed against the side of the cube to move it in any direction. Thirty seven percent of participant pairs agreed on the referents *right* and *left*. We interpret this as meaning that when manipulating virtual objects, using direct manipulation techniques for translations is more natural.

However, when dealing with rotations, we saw more indirect manipulations in the form of circles made in the air around the axis of the intended rotation (Figure 4). A few people reached out and rotated the object directly (most common for roll, some occurrences in yaw). It is also of note that on a few occasions in the speech only block participants would make tracing gestures with their finger (Figure 4) for rotational referents. We speculate that this was to help lessen the cognitive challenge of figuring out which rotation was necessary by transferring the mental process to their visuospatial sketchpad. This follows previous findings that gestures help lighten the cognitive load of speech-based tasks [16].

During interviews after the experiment, 18/24 participants said that gestures were preferred for translations saying that gesturing took less thought. As seen in Figure 4 the most agreed upon proposals used reversible gestures for pairs of actions. This mirrors previous elicitation studies work [51, 64].

Select had the highest $\mathcal{AR}$ overall. This was due to the high occurrence of the legacy tap gesture (Figure 4). Legacy gestures were also produced for *Enlarge* and *Shrink*. Those being the two-finger zoom in /out from consumer touch screen phones. These gestures had a 12% co agreement rate. Meaning that while the gestures were highly agreed upon, pairs of participants were unlikely to agree on the same gestures for both referents. Legacy gestures are gestures that were used as inputs for previous technologies [41]. Legacy bias is viewed as negative when it does not utilize options available in the new input environment. This bias could be beneficial [8, 31, 46]. When appropriate for the new environment, legacy gestures provide the benefit of being more discoverable and more memorable to novice users [41].

6.2 Gesture and Speech Synergies

Delete and *create* had the lowest consensus in both the gestures and gesture and speech blocks. For these, speech might be the optimal input or a gesture derived by designers after doing a preference study. In the gesture and speech block, these referents had a higher CDR. Indicating that there was less disagreement between participants in the utterances proposed. Post-hoc analysis showed that while participants had less disagreement on the utterances proposed, they had higher disagreement on the appropriate gesture. Pointing to a location and saying "delete" or "remove" occurred with some frequency but the rate of snapping and blooming gestures lowered the overall AR (Figure 5). Even so, we believe that the benefit of improving the CDR makes these abstract commands well suited for co-occurring gesture and speech inputs. Other work has shown that producing gestures for abstract referents is difficult for some users, further bolstering this argument [51].

When only speech was allowed, most people use <action> <direction> or <action> <object> <direction> syntax such as "move left" or "move the cube left". When switching to multimodal inputs, people used more deictic gestures paired with an <action> <phrase> or action-gesture paired with a manipulation phrase. This is seen as a pointing gesture and saying "move" followed by a finger flicking in the direction of the intended movement.

Participants would use gestures to help with speech in the speech only block indicating a preference for multimodal interactions for rotational referents. Disfluent language (saying "left" when you mean "right") can be reduced by up to 50% when using multimodal gesture and speech [48]. This is due to the difficulties that most people have with spatial information, which in this study were the difficulties found when determining the correct direction for the rotations. The gesture portion of these commands was typically a finger trace indicating the

orientation of the rotation, which helps resolve the issue of finding the right language to execute the rotation. Five participants spontaneously gave degrees when presented with rotational referents. This added fine-grained turn control is a another compelling reason for enabling multimodal interactions for precise rotations.

An important finding of this study is that the median time between when a gesture starts and an utterance starts is 130 milliseconds. This finding can help researchers set up recognition windows for interactive multi-modal systems by indicating what lengths of time to wait between those input modes. Additionally, this finding helps bridge the work of linguistics [27, 36, 39] to human computer interaction. This shows that some of what is known about human to human communication extends into multi-modal interactions within AR environments. Similar findings have been seen for deictic gestures [23,25,33]. These findings presented here indicate that the timing windows for more generic manipulation gestures also conform to this pattern.

7 Design Guidelines

An optimal system would allow for unimodal gesture, unimodal speech, and multimodal gesture and speech interactions. While a large portion of users enjoy gesture and speech interactions [11, 19], some users still prefer unimodal interactions. For many things, direct manipulation should be available, particularly in the case of translations. For rotations, multimodal gesture and speech interactions should also be allowed. For every manipulation action, reversible interactions should be used. These could look like the gestures shown in Figure 4. With speech, this is more difficult but possible in some cases where a word has a clear opposite (i.e. "create" and "destroy"). With speech, it is important to also use aliasing as suggested in [41, 64]. For example, the combination "create' or "destroy" and "new" or "delete" covered nearly all proposals. A few times referents had very close ties for the most agreed-upon gesture. Aliasing would be beneficial here as well. For zoom in the legacy, two-finger zoom won but the pinch and expand were close in proposal frequency, for that case, both gestures should be available.

Nearly every participant in both the speech and the gesture with speech block proposed an utterance that was <action> <direction> or <action> <object> <direction> (Table 4). With this observation, we believe that a word spotting algorithm paired with aliasing certain commands together would be sufficient for most speech interaction tasks. This trend was also observed by [33].

When allowed to use both gesture and speech, gestures will typically proceed speech. Some of these gestures will be more generic pointing or turning gestures (screwing in a light bulb) accompanied by an action phrase such as "spin". Deictic gestures are more common when speech is allowed (Figure 5). The exception is that select had nearly all deictic gestures.

When developing a recognizer system for gesture and speech inputs the timing windows of co-occurring gestures and speech should be considered. When establishing time windows for speech centering the window around ˜130 milliseconds after gesture imitation would be beneficial. It should also be noted that each channel provides inputs that are disambiguated with the other channel. Seen in the pointing gesture paired with "delete" or "new" command.

The gestures proposed in this study can be implemented using the sensors built into consumer available AR-HMDs using either the stock hand tracking application program interface or the raw video stream. We found that tracking a few points (e.g., index tip, thumb tip, thumb base) was sufficient for direct manipulations and allowed for variations in the count of fingers used while gesturing. We recommend aliasing gestures across aliasing manipulative gestures across hand positions (open hand, pinch, grab, index only) based on the axis of movement. We also recommend that each one-handed interaction has a symmetric bi-manual version (i.e., one-hand push with two-hand push). While bi-manual gestures were not the most common interaction proposal in this study, other research suggests that with larger objects users opt for larger gestures [52, 57].

8 Limitations of the Study

While the findings presented here are important, this study has limitations. The environment presented uses one virtual object at a time. While this was by design, it is not clear if the findings will transfer into more complex environments (e.g., Lego-like applications) where object selection is necessary before a command it given. The design choice to ask the referent by using text in the virtual environment, while not uncommon, may have primed some of the participants' speech. Future work will address some of these limitations.

9 Conclusion

This is one of the first studies to test each of these input modalities independently in a within-subject design allowing us to take a more granular approach to the analysis of co-occurring gesture and speech usage within this environment. Due to that approach we are able to discuss the individual and joint strengths of each modality have been examined and suggestions have been made for both the unimodal and multimodal usage of these modes of interaction. This work extends the work of many linguists [27, 36, 39], and the work of computer scientists [4, 5, 12, 34] into AR-HMD building environments by examining the syntax patterns of co-occurring speech and gestures as compared to speech alone. We have shown that the timing between co-occurring manipulative gestures and speech in AR-HMD environments follows the same trend as found in studies using other types of gestures. This finding can be leveraged to create better recognizer systems as well as more natural human-centric interfaces. This study presents a set of user derived ego-centric gestures for use in AR building environments. These ego-centric gestures are critical when using a head-mounted camera such as the ones found on most AR devices. We have also found indications that gesturing is used to reduce cognitive effort when determining the direction of a requested rotation.

9.1 Future Work

Multiple unanswered questions require further work. For example, would the findings here translate to more complex environments? What if there are multiple users (either in the same room or not) in a shared virtual environment, would this lead to similar findings as human-to-human communications (e.g. [36]). Another future direction is to perform a follow-up study where the users are asked to generate gestures by seeing the movement of the object (with no text in the virtual environment). These questions are still open for any team to further explore. Head position and gaze were not measured in this study because there was only a single object presented at a time. In future work we plan to assess the role of gaze and head position in multi-object environments. Both gaze and head position serve as passive inputs that can improve accuracy in selection and interaction tasks.

10 Acknowledgements

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] B. Altakrouri, D. Burmeister, D. Boldt, and A. Schrader. Insights on the impact of physical impairments in full-body motion gesture elicitation studies. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 5:1–5:10. ACM, New York, NY, USA, 2016. doi: 10.1145/2971485.2971502

[2] D. Anastasiou, C. Jian, and D. Zhekova. Speech and gesture interaction in an ambient assisted living lab. In *Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments*, SMIAE '12, pp. 18–27. Association for Computational Linguistics, Stroudsburg, PA, USA, 2012.

[3] M. Z. Baig and M. Kavakli. Qualitative analysis of a multimodal interface system using speech/gesture. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 2811–2816, May 2018. doi: 10.1109/ICIEA.2018.8398188

[4] R. A. Bolt. “put-that-there”: Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.*, 14(3):262–270, July 1980. doi: 10.1145/965105.807503

[5] M.-L. Bourguet and A. Ando. Synchronization of speech and hand gestures during multimodal human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pp. 241–242. ACM, Apr. 1998.

[6] S. Buchanan, B. Floyd, W. Holderness, and J. J. LaViola. Towards user-defined multi-touch gestures for 3D objects. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pp. 231–240. ACM, Oct. 2013.

[7] S. Carbini, L. Delphin-Poulat, L. Perron, and J.-E. Viallet. From a wizard of oz experiment to a real time speech and gesture multimodal interface. *Signal Processing*, 86(12):3559–3577, 2006.

[8] E. Chan, T. Seyed, W. Stuerzlinger, X.-D. Yang, and F. Maurer. User elicitation on single-hand microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3403–3414. ACM, May 2016.

[9] A. Cohé and M. Hachet. Understanding user gestures for manipulating 3d objects from touchscreen inputs. In *Proceedings of Graphics Interface 2012*, GI '12, pp. 157–164. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 2012.

[10] S. W. Cook and M. K. Tanenhaus. Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, 113(1):98–104, 2009.

[11] A. Corradini and P. R. Cohen. On the relationships among speech, gestures, and object manipulation in virtual environments: Initial evidence. In *Advances in Natural Multimodal Dialogue Systems*, pp. 97–112. Springer, 2005.

[12] A. Corradini and P. R. Cohen. On the relationships among speech, gestures, and object manipulation in virtual environments: Initial evidence, 2005.

[13] A. Dünser, R. Grasset, H. Seichter, and M. Billinghurst. Applying hci principles to ar systems design. 2007.

[14] J. Eisenstein and C. M. Christoudias. A salience-based approach to gesture-speech alignment. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 2004.

[15] S. Goldin-Meadow, M. W. Alibali, and R. B. Church. Transitions in concept acquisition: using the hand to read the mind. *Psychological review*, 100(2):279, 1993.

[16] S. Goldin-Meadow, H. Nusbaum, S. D. Kelly, and S. Wagner. Explaining math: gesturing lightens the load. *Psychol. Sci.*, 12(6):516–522, Nov. 2001.

[17] A. Gupta, T. Pietrzak, C. Yau, N. Roussel, and R. Balakrishnan. Summon and select: Rapid interaction with interface controls in mid-air. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, pp. 52–61. ACM, New York, NY, USA, 2017. doi: 10.1145/3132272.3134120

[18] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139 – 183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9

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

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

[21] A. G. Hauptmann and P. McAvinney. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*, 38(2):231 – 249, 1993. doi: 10.1006/imms.1993.1011

[22] L. Hoff, E. Hornecker, and S. Bertel. Modifying gesture elicitation: Do kinaesthetic priming and increased production reduce legacy bias? In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 86–91. ACM, Feb. 2016.

[23] S. Irawati, S. Green, M. Billinghurst, A. Duenser, and H. Ko. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In *International Conference on Artificial Reality and Telexistence*, pp. 272–283. Springer, 2006.

[24] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. Unification-based multimodal integration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 281–288. Association for Computational Linguistics, 1997.

[25] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th international conference on Multimodal interfaces*, pp. 12–19, 2003.

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

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

[28] S. Khan and B. Tunçer. Gesture and speech elicitation for 3d cad modeling in conceptual design. *Automation in Construction*, 106:102847, 2019.

[29] K. Kin, M. Agrawala, and T. DeRose. Determining the benefits of direct-touch, bimanual, and multifinger input on a multitouch workstation. In *Proceedings of Graphics interface 2009*, pp. 119–124. Canadian Information Processing Society, 2009.

[30] D. B. Koons, C. J. Sparrell, and others. Integrating simultaneous input from speech, gaze, and hand gestures. *MIT Press: Menlo Park, CA*, 1993.

[31] A. Köpsel and N. Bubalo. Benefiting from legacy bias. *interactions*, 22(5):44–47, Aug. 2015. doi: 10.1145/2803169

[32] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[33] M. Lee and M. Billinghurst. A wizard of oz study for an ar multimodal interface. In *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 249–256, 2008.

[34] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Real.*, 17(4):293–305, Nov. 2013.

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

[36] D. McNeill. *Gesture and Thought*. 2005.

[37] M. Micire, M. Desai, A. Courtemanche, K. M. Tsui, and H. A. Yanco. 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*, ITS '09, pp. 41–48. ACM, New York, NY, USA, 2009. doi: 10.1145/1731903.1731912

[38] C. Mignot, C. Valot, and N. Carbonell. An experimental study of future "natural" multimodal human-computer interaction. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pp. 67–68, 1993.

[39] L. Mol and S. Kita. Gesture structure affects syntactic structure in speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, 2012.

[40] M. R. Morris. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, pp. 95–104. ACM, New York, NY, USA, 2012. doi: 10.1145/2396636.2396651

[41] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, M. c. Schraefel, and J. O. Wobbrock. Reducing legacy bias in gesture elicitation studies. *Interactions*, 21(3):40–45, May 2014.

[42] M. R. Morris, J. O. Wobbrock, and A. D. Wilson. Understanding users' preferences for surface gestures. In *Proceedings of graphics interface 2010*, pp. 261–268. Canadian Information Processing Society, 2010.

[43] T. Moscovich and J. F. Hughes. Indirect mappings of multi-touch input using one and two hands. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1275–1284. ACM, 2008.

[44] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson. Memorability of pre-designed and user-defined gesture sets, 2013.

[45] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*, pp. 409–420. Springer Berlin Heidelberg, 2004.

[46] F. R. Ortega, A. Galvan, K. Tarre, A. Barreto, N. Rishe, J. Bernal, R. Balcazar, and J. Thomas. 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)*, pp. 144–153, Mar. 2017.

[47] F. R. Ortega, K. Tarre, M. Kress, A. S. Williams, A. B. Barreto, and N. D. Rishe. Selection and manipulation whole-body gesture elicitation study in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1723–1728, March 2019. doi: 10.1109/VR.2019.8798105

[48] S. Oviatt. Mulitmodal interactive maps: Designing for human performance. *Human–Computer Interaction*, 12(1-2):93–129, 1997.

[49] S. Oviatt. Taming recognition errors with a multimodal interface. *Communications of the ACM*, 43(9):45–51, 2000.

[50] S. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pp. 415–422, 1997.

[51] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-Defined gestures for augmented reality. In *Human-Computer Interaction – INTERACT 2013*, pp. 282–299. Springer Berlin Heidelberg, 2013.

[52] T. Plank, H.-C. Jetter, R. Rädle, C. N. Klokmose, T. Luger, and H. Reiterer. 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*, CHI '17, pp. 4548–4560. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025537

[53] S. Robbe. 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*, pp. 349–350, 1998.

[54] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction, 2011.

[55] J. Ruiz and D. Vogel. Soft-Constraints to reduce legacy and performance bias to elicit whole-body gestures with low arm fatigue. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3347–3350. ACM, Apr. 2015.

[56] E. A. Schegloff. On some gestures' relation to talk.(pp. 266-296) in j. maxwell and j. heritage (eds.) structures of social action, 1984.

[57] K. Tarre, A. S. Williams, L. Borges, N. D. Rishe, A. B. Barreto, and F. R. Ortega. 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*, VRST '18, pp. 125:1–125:2. ACM, New York, NY, USA, 2018. doi: 10.1145/3281505.3281590

[58] T. Tsandilas. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Trans. Comput. Hum. Interact.*, 25(3):18, June 2018.

[59] R.-D. Vatavu and J. O. Wobbrock. 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*, pp. 1325–1334. ACM, Apr. 2015.

[60] R.-D. Vatavu and J. O. Wobbrock. Between-Subjects elicitation studies: Formalization and tool support. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3390–3402. ACM, May 2016.

[61] P. Vogiatzidakis and P. Koutsabasis. Gesture elicitation studies for Mid-Air interaction: A review. *Multimodal Technologies and Interaction*, 2(4):65, Sept. 2018.

[62] M. L. Wittorf and M. R. Jakobsen. Eliciting Mid-Air gestures for Wall-Display interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 3:1–3:4. ACM, New York, NY, USA, 2016.

[63] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers. Maximizing the guessability of symbolic input, 2005.

[64] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009.

[65] K. Wolf, A. Naumann, M. Rohs, and J. Müller. 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*, INTERACT'11, pp. 559–575. Springer-Verlag, Berlin, Heidelberg, 2011.

[66] I.-A. Zaiţi, Ş.-G. Pentiuc, and R.-D. Vatavu. On free-hand TV control: experimental results on user-elicited gestures with leap motion. *Pers. Ubiquit. Comput.*, 19(5):821–838, Aug. 2015.