It's a Joint Effort: Understanding Speech and Gesture in Collaborative Tasks

Isaac Wang[1](✉), Pradyumna Narayana[2], Dhruva Patil[2], Rahul Bangar[2], Bruce Draper[2], Ross Beveridge[2], and Jaime Ruiz[1]

[1] Department of CISE, University of Florida, Gainesville, FL 32611, USA
wangi@ufl.edu
[2] Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA

Abstract. Computers are evolving from computational tools to collaborative agents through the emergence of natural, speech-driven interfaces. However, relying on speech alone is a limitation; gesture and other non-verbal aspects of communication also play a vital role in natural human discourse. To understand the use of gesture in human communication, we conducted a study to explore how people use gesture and speech to communicate when solving collaborative tasks. We asked 30 pairs of people to build structures out of blocks, limiting their communication to either *Gesture Only*, *Speech Only*, or *Gesture and Speech*. We found differences in how gesture and speech were used to communicate across the three conditions and found that pairs in the *Gesture and Speech* condition completed tasks faster than those in *Speech Only*. From our results, we draw conclusions about how our work impacts the design of collaborative systems and virtual agents that support gesture.

Keywords: Gesture · Speech · Multimodal · Communication · Collaboration

1 Introduction

Interaction with computers has become increasingly natural through the emergence and near-ubiquity of speech-driven interfaces, such as Apple's Siri [1] and Amazon's Alexa [2]. These dialogue-based systems have enabled computers to become more like collaborative agents rather than computational tools. Any user can now talk and interact with these systems much as they would with another human, making interactions more natural.

However, these interfaces lack the richness and multimodality of human-to-human communication. Our everyday discourse encompasses much more than just speech; throughout the course of a conversation, we may use gesture, expression, body language, and/or context to communicate our feelings, thoughts, and actions [3–7]. Thus, the use of speech as the sole medium of communication with speech-driven interfaces is a limitation. Computers must utilize nonverbal channels of communication if they are to function effectively as collaborative partners by emulating human interactions.

Thus, we need to identify and model how gestures and speech are used in human-to-human communication. Prior work has studied gesture interaction quite extensively

M. Kurosu (Ed.): HCII 2021, LNCS 12763, pp. 159–178, 2021.
https://doi.org/10.1007/978-3-030-78465-2_13

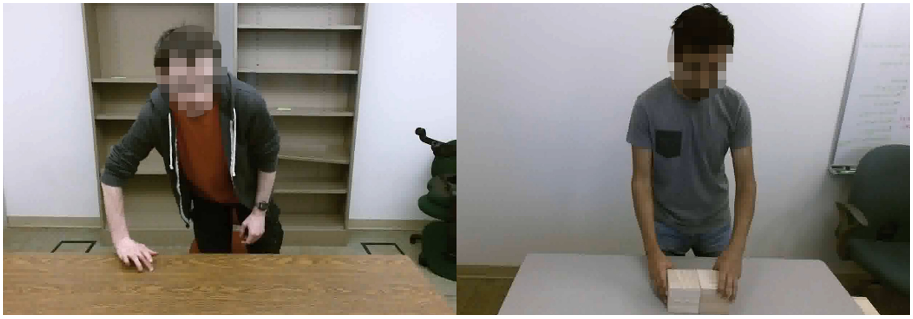


Fig. 1. Example showing two people collaborating to build a structure. The Signaler (left) is gesturing to the Actor (right).

(e.g., [8–11]), and has also investigated the use of multimodal interaction (e.g., [12–15]). Other studies have focused on understanding how to support remote collaboration between people by supporting the coordination of mutual knowledge through common ground [16–19]. These studies have mainly concentrated on the usefulness of shared visual information [18, 20, 21] and nonverbal information other than gesture, such as gaze [22–24]. Our goal, however, is to motivate and inform the development of multimodal interfaces and collaborative agents that incorporate the use of gesture. We work toward this goal by studying the natural interaction between two people working to solve a collaborative task (Fig. 1).

In this paper, we present an exploratory study observing the natural dyadic communication between two people collaborating to build structures out of blocks. We limit the communication allowed across three conditions: *Gesture Only*, *Speech Only*, and *Gesture and Speech*. Our work provides the following main contributions:

1. We describe *differences in the usage of speech and gesture* when the two are used together, or when one is present and the other is not.
2. We show how *communication was predominantly multimodal* when people were given the option to use both gesture and speech.
3. We present data showing a *quantifiable performance benefit from the inclusion of gesture with speech.*
4. From our findings, we derive *insights and design implications* into how the inclusion of gesture can benefit collaboration and the design of collaborative virtual agents.

These contributions highlight the importance of integrating the use of speech and gesture when designing collaborative technology and collaborative virtual agents.

2 Related Work

There has been an ongoing effort in HCI on integrating both speech and gesture in multimodal computer interaction since the introduction of the point-and-manipulate "Put-that-there" system by Bolt [25]. In this section, we focus on prior gesture research that motivates our work. This includes research on gesture in the domains of psychology and HCI as well as research on multimodal communication and collaboration.

2.1 Gesture in Psychology

Research in multimodal gesture and speech interaction began with psychological and cognitive studies (e.g., [3, 7, 26, 27]). For instance, Kendon's gesture classification scheme [7, 27] describes five categories of gestures that range from gesticulation, which appears in the presence of speech, to sign languages, which appear in the complete absence of speech. Additionally, McNeill [7] emphasized that gesture and speech are closely intertwined and both are critical to organization and generation of thought. Other work by Kendon details how gesture is used in everyday conversation to refer to objects and to add expressiveness to language by demonstrating events or actions [6]. These gestures co-occur with speech and may contain redundant or complementary information. Both Kendon and McNeill emphasize that gesture contains information. Therefore, it is likely that understanding both speech and gesture can help a system better interpret what a person is trying to convey.

Kendon and McNeill focused on the descriptive use of gesture in the context of narration and conversation. Gestures also play a more functional role in communication, such as conveying spatial information [6, 28, 29]. Gestures and other nonverbal cues are also valuable in conversational grounding, i.e. the process of establishing a mutual understanding in communication [4, 5, 30]. The use of gesture also helps manage the flow of a conversation by functioning as a signal for turn-taking and providing a "back-channel" for communicating attention and turn-taking [26, 31]. Additionally, Argyle [3] noted how gestures have social functions, such as their use in simple greetings or their conveyance of emotional state.

The concepts and theories derived from these studies help motivate our work in designing systems that incorporate the use of gesture. These studies illustrate how gesture is an important and natural part of human communication, both in interacting with other humans and in helping to organize and express thoughts. Therefore, designing systems that recognize and understand gesture is vital if we wish to create collaborative technology and virtual agents that are more natural and versatile.

2.2 Gesture in HCI

In HCI, there has been a large effort in creating systems focused on gesture interaction. There is a large body of prior work on designing and developing gesture input techniques and recognition, ranging from mobile devices [8, 10, 32, 33] to large-screen displays [11, 34, 35], as well as research on multimodal interaction [12–15, 36, 37]. This body of work focuses on using gesture as an alternative means of providing input commands to a system.

In contrast, other work focuses on the use of gesture and other nonverbal expressions in a more communicative manner, emphasizing how gesture can also be used to convey information. For instance, Grandhi et al. [38] saw how people used pantomime to directly paint images of objects and actions (i.e., holding imaginary objects and pretending to perform a task) rather than attempt to use abstract gestures to describe the task or object.

Holz and Wilson [9] showed that pantomimes can describe the specific spatial features of 3D objects. Likewise, Sowa and Wachsmuth [39] detailed a method to allow users to select objects using gestures by inferring this spatial information from their hands.

Gesture can also convey information alongside speech. Epps et al. [40] conducted a study on multimodal expressions for photo manipulation tasks, and found that gestures provided complementary information (e.g., pointing to an object and giving a verbal command) and redundant information (e.g., a rotational gesture along with speech describing the rotation). Furthermore, other studies [41–43] also noted and explored the varying co-expressiveness and redundancy of gesture and speech. Our work extends these studies by similarly looking at how speech and gesture are used together to communicate and convey information, but also compares how using gesture and speech separately differs from a multimodal usage.

2.3 Multimodal Communication and Collaboration

As our work focuses on the use of speech and gesture in the context of collaboration, we not only draw from prior work on gesture and multimodal interaction but also work on collaboration and computer-mediated communication. For instance, Bekker et al. [44] conducted an analysis of gestures in a face-to-face design session and found that gestures are naturally used to describe actions, depict spatial relations, and refer to specific items. They also found that gestures help manage the flow of conversations, similar to Duncan and Niederehe's [26] results that showed how people use gestures to signal when they want to speak. Similarly, in observing interactions between people across video, Isaacs and Tang [45] pointed out that video helps express and enhance understanding through the conveyance of nonverbal information, such as facial expression and occasionally gesture. Others have looked at shared gaze information to help resolve references when collaborating [22–24].

Several studies also emphasize the value of providing shared visual information (such as a shared workspace) when supporting virtual collaboration between people [18, 20, 21, 46]. These studies hold that allowing people to share a view of task objects and actions performed on those objects helps with grounding and is thus critical to collaboration. Building off of these studies, Fussell et al. [16] explored ways of using "embodiments of gesture," or the use of alternatives such as pointers or pen drawings in place of natural gestures through body motions. Similarly, Kirk et al. [17] showed how allowing people to gesture at objects by projecting their hands onto the workspace can also help facilitate grounding.

We seek to extend these studies by understanding how the modalities of speech and gesture are used in face-to-face interactions. We recognize the benefit of a shared workspace, and therefore include a shared view in all conditions. Our goal is then to understand and show the added benefit of gesture *in addition to* the shared workspace. To our knowledge, no prior work has investigated how gesture and speech are used separately and together when supporting a natural face-to-face interaction in the context of a shared visual workspace.

3 Method

We conducted an exploratory study asking pairs of adults to collaboratively build different pre-determined structures using wooden blocks.

3.1 Participants

We recruited 60 participants (10 pairs per condition) from computer science classes at a local university and through word-of-mouth recruiting. Participants were between the ages of 19 and 64 (mean = 24.2, SD = 7.7), and 17 participants were female. Out of the 60 participants, six were left-handed, and one person was ambidextrous. Half of all participants had prior experience with motion gesture systems such as the Microsoft Kinect. Out of the 30 pairs of participants, 16 pairs were previously acquainted with each other. All participants received a $10 Amazon gift card as compensation. Our study was approved by our Institutional Review Board.

3.2 Procedure

Each pair of participants was randomly assigned to one of the experimental conditions at the start of the study. They were then split up into the separate rooms, each with a computer and a display that allowed them to communicate. One participant was assigned the role of *Actor*, and the other, the role of *Signaler*. For each trial, the Signaler was given a picture of a *block layout*, depicting an arrangement of blocks that represented the end goal. The Actor was provided a set of blocks. The task was for the Signaler to direct the Actor to arrange the blocks to match the goal layout.

The pair was asked to complete two sessions. For each session, they were asked to complete up to 10 tasks within 30 min. We ended the session if time expired. After the first session, the participants switched roles and attempted another session for a total of up to 20 tasks. The tasks used for each pair were randomly selected without replacement from a corpus of 80 unique layouts.

Since we wanted to observe natural communication in action, participants were not allowed to talk or strategize beforehand and were given no instruction on how to speak or gesture. A trial began when the experimenter presented a new block layout to the Signaler and ended when the Actor replicated the block layout. The time taken to complete each task was measured, and video of both the Actor and Signaler was recorded for later analysis.

3.3 Apparatus

The Signaler and Actor both interacted with similar systems during the study. Each participant stood in front of a table facing a TV screen on the opposite end of the table (Fig. 2). A Microsoft Kinect v2 sensor [47] was also set up on the opposite end, facing the participant. These were connected to a computer that drove the display and collected data from the Kinect. We developed software to stream live video (and audio) from the Kinect sensors between the two setups. High-quality video (1080p) was streamed at

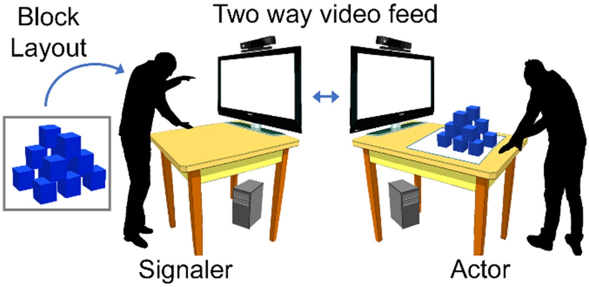


Fig. 2. The experiment setup. The participant on the left (Signaler) was asked to direct the other participant (Actor) to construct a block layout.

30 FPS with no noticeable latency. Because the Kinect sensors capture a mirror image (horizontally reflected; left appears as right, and vice versa), we corrected the image when streaming between the two setups. The Kinect sensors were also used to record RGB video, depth data, and motion capture skeletons.

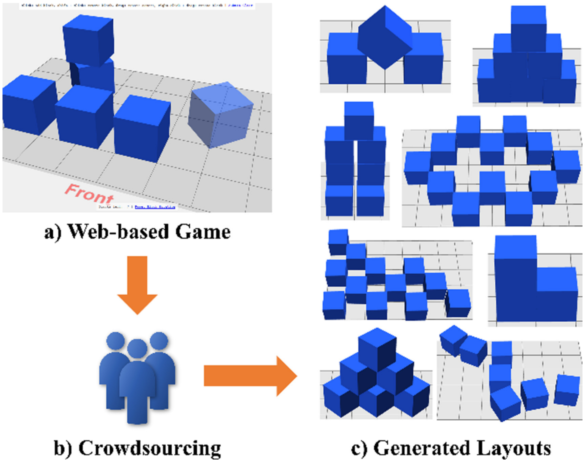
This full-duplex video link allowed the Signaler and Actor to communicate with each other as if they were facing each other at opposite ends of the same table. The table acted as a shared workspace, as blocks placed on the table could be seen by both participants (from opposite perspectives).

The blocks used in the study were wood cubes with 4-in. sides (10.6 cm). The Actor was provided with 12 blocks total; however, not all 12 blocks were used in every layout. These blocks were kept off the table and placed on the Actor's right-hand side. Blocks were removed from the table and reset between trials. Participants stood approximately 5 feet (1.5 m) away from the screen, allowing for a visible workspace of around 5 × 2.5 feet (1.5 × 0.76 m) on the table.

3.4 Block Layouts

We used a set of 80 different block layouts for the study. Crowdsourcing was used to quickly collect a large variety of layouts that would also afford different intents. For instance, an arrangement of blocks with multiple layers may involve gestures or speech to stack blocks on top of each other, and an arrangement with rotated blocks would involve gestures or speech specifically to rotate blocks. Thus, using a crowdsourced collection of layouts in our study allowed us to capture gestures and utterances for a wide range of possible intents.

To collect layouts, we developed and deployed a web-based game in which submitted layouts were scored based on how closely the position and rotation of blocks matched those in previously submitted layouts. The game simulated the arrangement of blocks in a virtual 3D environment (Fig. 3a). Block physics were also simulated, allowing blocks to support each other and lean against each other in different orientations, further opening possibilities for creative arrangements. We distributed the game to local computer science classes, asking students to submit their own layouts and compete for the high score. We collected a corpus of over 200 layouts in three weeks. Out of these, we first excluded layouts that were unstable (or fell over) or were not unique, and then randomly selected



a) Web-based Game

b) Crowdsourcing

c) Generated Layouts

Fig. 3. We (a) deployed a web-based blocks game to quickly (b) crowdsource a diverse corpus of (c) block layouts.

a set of 80 distinct layouts for this experiment. The layouts used varied in their difficulty and complexity (Fig. 3c).

3.5 Conditions

We restricted the participants' communication modalities based on the following conditions:

- *Gesture Only:* Participants could only see each other using the video presented on the TVs; the audio was muted, restricting participants to using only non-verbal communication to accomplish tasks.
- *Speech Only:* Participants were not able to see each other and could only rely on speech to communicate; the video feed was restricted so that both participants could only see the blocks on the table. This was done to remove the ability to communicate through gestures while still retaining a shared workspace.
- *Gesture and Speech:* Both audio and video were enabled, allowing participants to use any natural combination of speech and gesture they wished.

3.6 Data Analysis

To compare gesture use between the *Gesture Only* and *Gesture and Speech* conditions, we annotated the video recordings to label the occurrences of different gestures. Our focus was on how required actions and specific features of the layout were communicated. Therefore, we focused on the gestures made by the Signaler. We followed the gesture analysis procedure detailed by Wang et al. [48]. Using the video annotation tool described in their paper, we marked the start and end of each observed gesture and assigned it a label that described the physical motion of body parts that were involved in making the gesture.

A specific labeling language was used to keep labeling consistent and descriptive. See [48] for further details.

Additionally, the implied intent of each gesture was annotated, e.g., a gesture was given the intent of "slide left" if it signified that the current block(s) needed to be moved left. Intents included labels for sliding/moving blocks, stacking, signifying "OK," "stop," etc. Labels used for gesture and intent were kept consistent across the two conditions, to facilitate comparing instances of gestures between them.

We generated speech transcripts from the videos in the *Speech Only* and *Gesture and Speech* conditions in order to analyze speech. We used an automatic speech recognition (ASR) service by IBM Watson [49] to retrieve transcripts as well as timing information for the beginning and end of utterances. We also ran the videos through Google Cloud Speech [50] to compare ASR accuracies and found the accuracies to be similar to Watson.

4 Results

Using both our labeled gesture data and generated transcripts, we analyzed how speech and gesture were used in the different conditions, and whether the modality used in each condition influenced task performance.

4.1 Differences in Gesture Use

We labeled 24,503 gesture instances across all trials. Additionally, we note that there were 15,222 (62.1%) gestures in the *Gesture Only* condition and only 9,281 (37.9%) in the *Gesture and Speech* condition. Our dataset had 5,060 unique gesture labels and 187 unique intent labels. However, due to the nature of our gesture labels as compound physical descriptions, many of these only occurred once in the entire set. Only 1,427 gestures occurred two or more times, and only 110 gestures occurred at least twenty times and were used by more than one participant.

Due to the large number of unique gestures and intents (caused by the level of detail in our labels), we preprocessed the data by filtering, grouping, and categorizing gestures to decrease the number of items to a manageable set. This was done to extract themes to compare gesture use between the *Gesture Only* and *Gesture and Speech* conditions.

We first filtered the list of unique gestures and intents to include only those that were performed by at least four people and were observed to have occurred in at least 20 instances across the entire set. This allowed us to take a high-level look at the gestures and intents that were commonly used. Similar gestures and intents were grouped into 25 discrete actions and organized into four types: Numeric (counting gestures), Command (translate, rotate, etc.), Reference (this block, there, here, etc.), and Social Cues (start, done, OK, no, stop, etc.). It is important to note that we also included the absence of action ("wait" and "think") under Social Cues, to also look at potential differences in how the Signaler paused to think or waited for the Actor to complete an action. For each action, we compared the number of times the action occurred in the *Gesture Only* condition against the number of times it occurred in the *Gesture and Speech* condition. These frequencies are presented in Table 1.

Table 1. Comparison of action frequencies between the *Gesture Only* and *Gesture and Speech* conditions. Shaded actions occur at least four times more in *Gesture Only*

	Action	Gesture Only	Gesture & Speech
Numeric	one	85	18
	two	64	9
	three	26	4
	four	21	5
	five	29	8
Command	separate	126	125
	translate	234	73
	continual translate	206	44
	translate towards	35	20
	together	83	38
	rotate	225	67
Reference	that / there	150	106
	here / this	111	42
	this group	86	28
	this block	240	94
	these blocks	24	20
	this stack	38	16
	this column	52	13
Social Cues	start	100	51
	ok	693	225
	wait	1582	1555
	think	400	246
	done	81	47
	no	109	11
	emphasis	17	27

We discovered that seven actions were used at least four times more often in the *Gesture Only* condition than in *Gesture and Speech* (highlighted in Table 1). Four of them were the numeric gestures one through four, so these were grouped together, giving us four distinct groups:

- Numeric gestures ("one" through "four"): referring to a quantity by holding up the number of fingers.
- "Continual translate": the act of continually translating an object in a direction until feedback (as in a stop or OK sign) is given.
- "This column": referring to a series of blocks arranged outwardly, as opposed to vertically or horizontally.
- "No": a Social Cue used to give negative feedback.

4.2 Differences in Speech Use

We analyzed speech use between the *Speech Only* and *Gesture and Speech* conditions with the goal of identifying high-level themes by looking for patterns in common phrases people used. Due to the nature of speech, many words or phrases only appeared once or twice, impeding direct comparison between the two conditions. Thus, we looked at the most common bigrams and trigrams retrieved from the transcripts and compared their frequencies. Results from this analysis showed that the most common bigrams described spatial relations through prepositional phrases such as "in the" or "on top." We also looked for items that appeared at least four times more often in either condition, similar to our analysis of gestures. Given these criteria, we did not find any major differences in speech use between the two conditions. There may be additional information gained from looking at the speech acts used or how turn-taking was accomplished, but any deeper linguistic analysis is beyond the scope of this paper.

When comparing the occurrences of trigrams, we gain more insight into how speech was used. We saw the use of prepositional phrases to describe where to place blocks in relation to one another. This paints a clear picture of their use, with phrases such as "in the middle" and "on top of." We also see a large use of the phrases "a little bit" and "little bit more," used to describe small movements or continuations of movements such as "*now go forward a little bit*" (P15) and "*little bit more up and then pushed in a little bit*" (P29).

When examining the top 20 trigrams, we find only a few that appeared much more often in the *Speech Only* condition than *Gesture and Speech*. The trigram "there you go" appears 7.5 times more often in *Speech Only*; however, upon further analysis, this was due to a participant (P57) inflating the occurrence count by repeatedly using the phrase in a trial. The trigrams "going to be" and "little bit more" were used five times and four times more often, respectively. The larger occurrence of "going to be" (e.g., "*it's going to be on your left side*" – P27) is not particularly meaningful, and may just be a nuance of speech (this was not caused by an inflated count, unlike "there you go"). For "little bit more," we note that it is related to "a little bit," so possibly there were more cases where the Signaler needed to make finer adjustments to achieve the goal. We also note the larger occurrence of "I want you" and "want you to" in the *Gesture and Speech* condition. This was also due to a participant (P41) constantly repeating the phrase and inflating the count.

4.3 Co-expression of Speech and Gesture

We took a closer look at the *Gesture and Speech* condition, to identify how the two modalities were used together. Using both our labeled gesture data and transcribed utterance timings, we conducted an analysis following the methodology used by Oviatt and colleagues [15, 36, 40], correlating instances of gesture with specific utterances. This was accomplished by comparing the instances in which gestures occurred with the times where utterances occurred and counting the instances where they overlapped.

We observed a total of 8,473 multimodal and unimodal constructions. Of these, 66.5% (5,638 constructions) were multimodal and 33.5% (2,835) were unimodal. Of the unimodal constructions, 43.2% (1,226) used speech alone and 56.8% (1,609) used gesture alone (Fig. 4). These results mirror those detailed by Epps et al. [40], with the exception that we saw a larger number of gesture only constructions than speech only constructions.

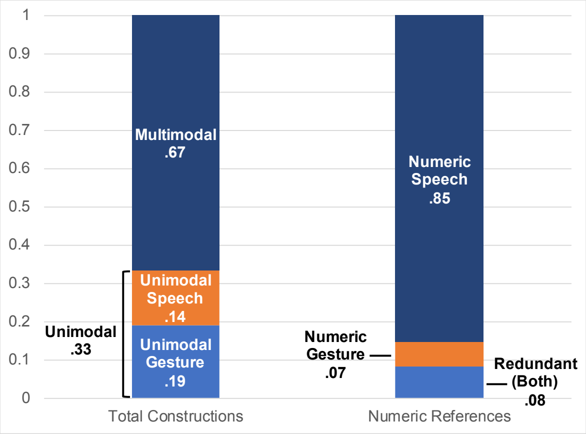


Fig. 4. Proportion of total multimodal/unimodal constructions (left) and numeric references (right) in the *Gesture and Speech* condition.

Our results on gesture use showed that the participants in the *Gesture and Speech* condition used fewer numeric gestures. To determine if speech was being used instead of gestures to refer to numbers, we took a closer look to see if one modality was used more than the other for expressing numbers. The right-most barchart in Fig. 4 depicts the proportions of numeric references using speech, gesture, or both. First, we saw that there was a total of 946 different times where the Signaler referred to a number (849 multimodal constructions, 97 unimodal), either for specifying the number of blocks to add or referring to a group of blocks already on the table. Out of these, an overwhelming majority (807, 85.3%) relied on speech alone to refer to numbers (no numeric gestures were observed concurrently), as opposed to 61 (6.4%) instances that relied on gesture alone to specify a number. There were also a few instances where gesture and speech were used redundantly, representing 78 (8.3%) of the constructions. These instances represented cases where participants verbally stated a number while also holding up the corresponding number of fingers. However, the majority of constructions relied on speech rather than gesture.

4.4 Task Performance

We compared task completion times to determine whether modality had an impact on task performance. As the number of trials differed between participants, task performance was evaluated by taking the average trial/task completion time for each session, resulting in two average times per pair (one for each Signaler/Actor assignment). We found one outlier (greater than three standard deviations away from the mean) in the *Gesture Only* condition. This was due to one pair (P39 & P40) having extreme difficulty with the task, resulting in an average task completion time that was eight standard deviations above the mean (299.1 s). Completion times for this pair were removed from further analysis. For the *Gesture Only* condition, the average trial time for a participant was 89.1 s (SD = 26.7), for *Gesture and Speech*, the average was 69.6 s (SD = 33.5), and for *Speech Only*, the average was 97.4 s (SD = 39.5).

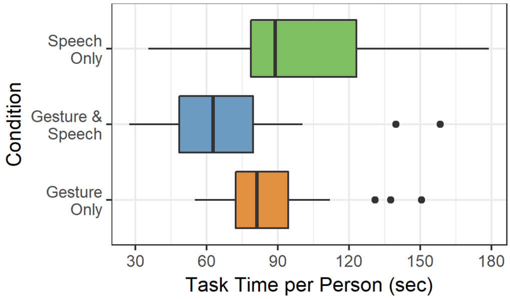


Fig. 5. Boxplots comparing average task time per person across conditions, after outlier removal. The Gesture and Speech condition was faster ($p < 0.05$) than Speech Only.

Analysis of variance (ANOVA) revealed a significant effect of condition on trial/task completion time ($F_{2,55} = 3.531, p < 0.05$). Post-hoc analysis using Bonferroni correction revealed that the *Gesture and Speech* condition was significantly faster than *Speech Only* ($p < 0.05$). However, we found no significant difference for completion time between *Gesture Only* and the two other conditions ($p > 0.25$ in both cases). Figure 5 shows the task completion times per person across conditions.

5 Discussion

We saw that participants predominantly communicated using both modalities simultaneously if given the option, as opposed to using either speech or gesture alone. Our results also suggest that speech use was the same in both speech conditions. Thus, we attribute the inclusion of gesture to the faster completion times found in the *Gesture and Speech* condition. We first discuss the differences in gesture and speech use and then focus on the ways that gesture improved communication between pairs.

5.1 Different Strategies for Different Modalities

From prior work [7, 51], we would expect that gesture use changes accordingly with the presence and absence of speech. This is in alignment with our results, which suggest that the use of gesture changes in the presence of speech. We interpret these findings as different communication strategies for different modalities. For instance, our results showed that numeric gestures were used less often in the *Gesture and Speech* condition. When communicating multimodally, speech was the dominant channel for communicating numbers. If numeric gestures were used, they were often redundant, with participants still using speech to specify numbers. It was likely easier to verbally state the number of blocks than perform a numeric gesture. Speech is also more powerful in this case because the Signaler could easily speak any number from one to twelve, but gesturing via finger counting has a limit of ten. This is supported by our data, which shows that every number from one to twelve was spoken at least once.

Similarly, it would likewise be easier or more understandable for the Signaler to say "no" or verbally correct the Actor than to shake their head or rapidly shake their hands back and forth to quickly stop or undo the current action (which is a drastic action in the videos, much like engaging the emergency brakes when the Actor performs an incorrect move). Furthermore, representing "this column" and "continual translate" through gestures may not be necessary because there are better words to describe the actions of specifying a column or continually translating an object. It is possible that gestures for "continual translate" were supplanted with phrases such as "a little" or "little bit more," avoiding the use of a continual movement in favor of fine-grained incremental movements.

When given the ability to use both speech and gesture, people used multimodal expressions two times more frequently than unimodal expressions. We observed that, when gesture was used with speech, it was often used to convey supplementary information as opposed to redundant information. We saw that gestures were used in the *Gesture and Speech* condition almost as often as in the *Gesture Only* condition. Signalers used these referential gestures (here, there, etc.) to depict or point to objects, along with speech to describe how to manipulate those objects. This is illustrated in Fig. 6, where the Signaler is pointing to a group of three blocks while saying, "Put *that* on the other side." In this case, gesture is used to resolve the verbal reference "that."

These results demonstrate how a multimodal system would need to adapt to support interaction in environments where one modality becomes unavailable. For example, if a user of a multimodal system is likely to encounter environments where speech becomes impractical (e.g., noisy environments), then the system must be prepared to recognize additional gestures that would not appear when speech is present.

5.2 Speech and Gesture for Resolving Orientation

We also saw how speech was used in conjunction with gesture to resolve orientation between pairs. When analyzing the videos, we noticed that people used several different reference frames and perspectives when communicating. For example, sometimes when the Signaler referred to "left" in either gesture or speech, they meant their own left. Other times, "left" referred to the Actor's left. Likewise, there was confusion regarding the "front" and "back" orientations, which could also refer to either the Signaler's perspective or the Actor's perspective. This was further compounded when front and back used a different perspective than left and right. For instance, front and back could refer to the Signaler's perspective while at the same time left and right would refer to the Actor's perspective.

Due to the different perspectives used, participants needed to first achieve common ground with respect to orientation. While we found that both gesture and speech could be ambiguous when conveying direction and orientation, possessives could be used in

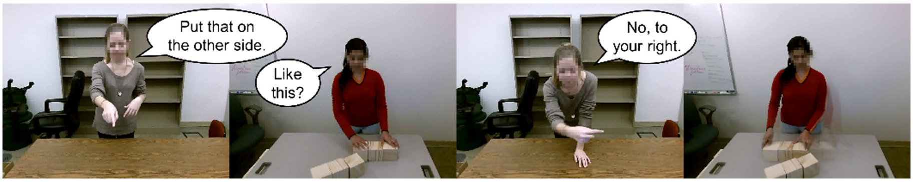


Fig. 6. In this example, the Signaler clarified orientation by saying "your right" while also pointing in the correct direction.

speech (e.g., "your" or "my") to quickly specify the perspective of the direction (e.g., "your left" or "my left"). In contrast, gesture lacks this precision and required pairs to resolve perspective and orientation first, which could be difficult to communicate with gesture alone. Ideally, speech can be used to augment gesture to clarify a precise orientation, helping the pair communicate quickly. Figure 6 illustrates an example where the Signaler used both speech and gesture to clarify an ambiguous direction.

Our work emphasizes that people use speech to resolve issues revolving around the use of different perspectives when giving directions, similar to observations by Schober [52]. While our work shows that resolving perspectives continues to be a challenge, it also demonstrates that the problem is more difficult than previously stated because users will assume different perspectives for different directions and orientations, requiring the use of speech and gesture to negotiate perspective. Only after resolving perspective could a person use the full spatial capabilities of gesture in conjunction with the precise nature of speech.

5.3 Improved Performance with Gesture

Our results showed that task completion time in the *Gesture and Speech* condition was significantly faster than *Speech Only*. People were naturally utilizing both speech and gesture to their advantage, when given the ability to use both channels of communication. This is evident by the fact that we saw that people used multimodal expressions more frequently than unimodal expressions, using gesture and speech together to resolve references and disambiguate orientation. Our results also suggest that there is information encoded in gesture and other nonverbal communication, which aligns with prior work [19, 41, 42].

Additionally, information can be conveyed through iconic gestures to visually show how to manipulate a block (such as translating or rotating) or through gestural social cues. These social cues included explicit cues such as "no" (e.g., shaking the head from side to side) and "ok" (a nod or thumbs up gesture), but also included subtler actions such as moving closer to the table/other person to signify "start" or moving away from the table to signify "done." The lack of motion or gesture can also convey information, as we saw the Signaler stop moving when thinking or waiting for the Actor to complete an action (the lack of motion likely indicated to the Actor that they "had the floor" and could perform an action or communicate back to the Signaler). Regardless of the exact information conveyed, we saw that including video of both people helped pairs complete tasks faster.

Our finding that pairs in the *Gesture and Speech* condition had faster task completion times than those in the *Speech Only* condition contradicts prior work. Although prior work [18, 20, 21, 53] states that gestures can be used efficiently to refer to task objects, they did not find any improved task performance. We attribute the observed benefit to two characteristics of our study:

1. Our study used a single video to view the shared workspace and the collaborator. Prior studies [20, 46, 54] required users to split their attention between a video of the shared workspace and a video of the collaborator.
2. Our study enabled participants to see each other in the context of the shared workspace. This mimicked a face-to-face interaction between people. This is in contrast with work by Fussell et al. [20, 46], which used video to mimic a side-by-side collaboration setup. In their setup, the camera and display placements did not mimic what each person would naturally see if they were situated in context.

Together, these characteristics allowed participants to identify their spatial relation to the blocks and the other participant, supporting the use of referential gestures to efficiently refer to objects on the table.

6 Implications for Design

Based on our results and discussion, we summarize the following recommendations for designing systems that support gesture interaction, multimodal communication, and virtual collaboration.

Design Virtual Agents to Utilize Human Gesture: Virtual agents should be designed to perceive and understand users' gestures. By taking advantage of the information encoded in a person's gestures, an agent can better understand human intent and function as a more effective communicator and collaborator. This may also include understanding social cues (e.g., subtle cues such as walking up to the table to begin interaction, or responding to acknowledgements such as ok/no). Such social cues are critical in communicating feedback before, during, and after a task. Although not as prominent as other gestures (such as deictic pointing), they are still useful and communicative.

Design Around Ambiguous Orientation: People may assume different perspectives when referring to objects, causing ambiguity when using speech or gesture to describe orientation. Systems that recognize either modality should be designed in a way to determine the user's intended frame of reference (through mechanisms such as looking for positive or negative feedback, or explicitly grounding by asking if this was the intended direction). Resolving orientation should be flexible and adapt to the user's mental model, whether references are from their perspective or another perspective. If the agent has a visual embodiment, the agent should be designed to use gesture in conjunction with speech to refer to objects in the task, providing additional information for humans to leverage.

Include Views of People, Situated in Context: When designing systems that support collaboration, ensure that all parties can fully see each other. Likewise, when designing a virtual agent to collaborate with humans, designers should situate the agent to mimic where a real person would stand or sit. This allows for the full multimodal use of gesture and other nonverbal communication, which can give people multiple ways for referring to and describing objects, helping them more effectively collaborate on tasks. It is also important to maintain a shared visual workspace for tasks, and, if possible, use views that enable people to see each other in the context of the workspace, mimicking how they would be able to see each other in real life.

Design for Different Modalities: Systems that use multimodal gesture and speech interaction should adapt to work even when limited to one modality by understanding that the use of gesture will change in the absence of speech, and thus the system will need to adapt to different communication strategies. This will ensure that an interaction system or virtual agent can help accomplish tasks even in suboptimal modalities and enable natural user interaction in diverse scenarios. For example, people are not always able to speak and/or gesture, such as when the hands/arms are occupied or when situated in a quiet/noisy environment where speech is not an option. Although we presented differences in gesturing strategies for a physical collaborative task, further work is necessary to understand the communication differences for all tasks and scenarios.

7 Conclusion

In this paper, we described the results of a study exploring how pairs of people use gesture and speech to communicate in collaborative physical tasks. We highlighted differences in gesturing strategies when used in the absence or presence of speech and looked at how gesture and speech were used multimodally. We showed that supporting the use of both gesture and speech by including views of both people resulted in faster task completion times. Additionally, we described how gesture and speech were used together to resolve ambiguity in expressing direction and orientation. From these results, we presented design implications for systems and virtual agents that support gestures, communication, and collaboration.

Acknowledgements. This work was partially funded by the U.S. Defense Advanced Research Projects Agency and the U.S. Army Research Office under contract #W911NF-15-1-0459.

References

1. iOS - Siri – Apple. http://www.apple.com/ios/siri/. Accessed 11 Jan 2017
2. Amazon Alexa. http://alexa.amazon.com/spa/index.html. Accessed 11 Jan 2017
3. Argyle, M.: Bodily Communication. Methuen, London; New York (1988)
4. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (eds.) Perspectives on Socially Shared Cognition, pp. 13–1991. American Psychological Association, Washington, DC, US (1991)

5. Clark, H.H., Wilkes-Gibbs, D.: Referring as a collaborative process. Cognition **22**, 1–39 (1986). https://doi.org/10.1016/0010-0277(86)90010-7
6. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, New York (2004)
7. McNeill, D.: Hand and Mind : What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992)
8. Harrison, C., Hudson, S.E.: Abracadabra: wireless, high-precision, and unpowered finger input for very small mobile devices. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 121–124. ACM, New York, NY, USA (2009). https://doi.org/10.1145/1622176.1622199
9. Holz, C., Wilson, A.: Data miming: inferring spatial object descriptions from human gesture. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 811–820. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1979060
10. Ruiz, J., Li, Y., Lank, E.: User-defined motion gestures for mobile interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 197–206. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1978971
11. Walter, R., Bailly, G., Valkanova, N., Müller, J.: Cuenesics: using mid-air gestures to select items on interactive public displays. In: Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services, pp. 299–308. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2628363.2628368
12. Brewster, S., Lumsden, J., Bell, M., Hall, M., Tasker, S.: Multimodal "Eyes-free" interaction techniques for wearable devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 473–480. ACM, New York, NY, USA (2003). https://doi.org/10.1145/642611.642694
13. Keates, S., Robinson, P.: The use of gestures in multimodal input. In: Proceedings of the Third International ACM Conference on Assistive Technologies, pp. 35–42. ACM, New York, NY, USA (1998). https://doi.org/10.1145/274497.274505
14. Madhvanath, S., Vennelakanti, R., Subramanian, A., Shekhawat, A., Dey, P., Rajan, A.: Designing multiuser multimodal gestural interactions for the living room. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 61–62. ACM, New York, NY, USA (2012). https://doi.org/10.1145/2388676.2388693.
15. Oviatt, S., DeAngeli, A., Kuhn, K.: 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. ACM, New York, NY, USA (1997). https://doi.org/10.1145/258549.258821
16. Fussell, S.R., Setlock, L.D., Yang, J., Ou, J., Mauer, E., Kramer, A.D.I.: Gestures over video streams to support remote collaboration on physical tasks. Hum. Comput. Interact. **19**, 273–309 (2004). https://doi.org/10.1207/s15327051hci1903_3
17. Kirk, D., Rodden, T., Fraser, D.S.: Turn it this way: grounding collaborative action with remote gestures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1039–1048. ACM, New York, NY, USA (2007). https://doi.org/10.1145/1240624.1240782
18. Kraut, R.E., Gergle, D., Fussell, S.R.: The use of visual information in shared visual spaces: informing the development of virtual co-presence. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 31–40. ACM, New York, NY, USA (2002). https://doi.org/10.1145/587078.587084
19. Veinott, E.S., Olson, J., Olson, G.M., Fu, X.: Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 302–309. ACM, New York, NY, USA (1999). https://doi.org/10.1145/302979.303067

20. Fussell, S.R., Kraut, R.E., Siegel, J.: Coordination of communication: effects of shared visual context on collaborative work. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 21–30. ACM, New York, NY, USA (2000). https://doi.org/10.1145/358916.358947
21. Gergle, D., Kraut, R.E., Fussell, S.R.: Action as language in a shared visual space. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 487–496. ACM, New York, NY, USA (2004). https://doi.org/10.1145/1031607.1031687
22. Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., Zelinsky, G.J.: Coordinating cognition: the costs and benefits of shared gaze during collaborative search. Cognition **106**, 1465–1477 (2008). https://doi.org/10.1016/j.cognition.2007.05.012
23. D'Angelo, S., Gergle, D.: Gazed and confused: understanding and designing shared gaze for remote collaboration. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2492–2496. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858499
24. Gergle, D., Clark, A.T.: See what I'M saying?: Using dyadic mobile eye tracking to study collaborative reference. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 435–444. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1958824.1958892
25. Bolt, R.A.: "Put-that-there": Voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270. ACM, New York, NY, USA (1980). https://doi.org/10.1145/800250.807503
26. Duncan, S., Niederehe, G.: On signalling that it's your turn to speak. J. Exp. Soc. Psychol. **10**, 234–247 (1974). https://doi.org/10.1016/0022-1031(74)90070-5
27. Kendon, A.: How gestures can become like words. In: Cross-Cultural Perspectives in Nonverbal Communication, pp. 131–141. Hogrefe, Toronto; Lewiston, NY (1988)
28. Alibali, M.W.: Gesture in spatial cognition: expressing, communicating, and thinking about spatial information. Spat. Cogn. Comput. **5**, 307–331 (2005). https://doi.org/10.1207/s15427633scc0504_2
29. Bergmann, K.: Verbal or visual? How information is distributed across speech and gesture in spatial dialog. In: Proceedings of Brandial 2006, the 10th Workshop on the Semantics and Pragmatics of Dialogue, pp. 90–97 (2006)
30. Dillenbourg, P., Traum, D.: Sharing solutions: persistence and grounding in multimodal collaborative problem solving. J. Learn. Sci. **15**, 121–151 (2006). https://doi.org/10.1207/s15327809jls1501_9
31. Young, R.F., Lee, J.: Identifying units in interaction: reactive tokens in Korean and English conversations. J. Socioling. **8**, 380–407 (2004). https://doi.org/10.1111/j.1467-9841.2004.00266.x
32. Butler, A., Izadi, S., Hodges, S.: SideSight: Multi-"Touch" interaction around small devices. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, pp. 201–204. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1449715.1449746
33. Kratz, S., Rohs, M.: Hoverflow: Exploring around-device interaction with ir distance sensors. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 42:1–42:4. ACM, New York, NY, USA (2009). https://doi.org/10.1145/1613858.1613912
34. Müller, J., Bailly, G., Bossuyt, T., Hillgren, N.: MirrorTouch: combining touch and mid-air gestures for public displays. In: Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services, pp. 319–328. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2628363.2628379

35. Walter, R., Bailly, G., Müller, J.: StrikeAPose: revealing mid-air gestures on public displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 841–850. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2470774
36. Oviatt, S., Coulston, R., Lunsford, R.: When do we interact multimodally?: Cognitive load and multimodal communication patterns. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 129–136. ACM, New York, NY, USA (2004). https://doi.org/10.1145/1027933.1027957
37. Voida, S., Podlaseck, M., Kjeldsen, R., Pinhanez, C.: A study on the manipulation of 2D objects in a projector/camera-based augmented reality environment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 611–620. ACM, New York, NY, USA (2005). https://doi.org/10.1145/1054972.1055056
38. Grandhi, S.A., Joue, G., Mittelberg, I.: Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 821–824. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1979061
39. Sowa, T., Wachsmuth, I.: Interpretation of Shape-related Iconic Gestures in Virtual Environments. In: Wachsmuth, I., Sowa, T. (eds.) GW 2001. LNCS (LNAI), vol. 2298, pp. 21–33. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47873-6_3
40. Epps, J., Oviatt, S., Chen, F.: Integration of speech and gesture inputs during multimodal interaction. In: Proceedings of the Australian Conference on Human-Computer Interaction (2004)
41. Pfeiffer, T.: Interaction between Speech and Gesture: Strategies for Pointing to Distant Objects. In: Efthimiou, E., Kouroupetroglou, G., Fotinea, S.-E. (eds.) GW 2011. LNCS (LNAI), vol. 7206, pp. 238–249. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34182-3_22
42. Quek, F., et al.: Multimodal human discourse: gesture and speech. ACM Trans. Comput-Hum Interact **9**(3), 171–193 (2002). https://doi.org/10.1145/568513.568514
43. Ruiz, N., Taib, R., Chen, F.: Examining the redundancy of multimodal input. In: Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments, pp. 389–392. ACM, New York, NY, USA (2006). https://doi.org/10.1145/1228175.1228254
44. Bekker, M.M., Olson, J.S., Olson, G.M.: Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. In: Proceedings of the 1st Conference on Designing Interactive Systems: Processes, Practices, Methods, & Techniques, pp. 157–166. ACM, New York, NY, USA (1995). https://doi.org/10.1145/225434.225452
45. Isaacs, E.A., Tang, J.C.: What video can and can't do for collaboration: a case study. In: Proceedings of the First ACM International Conference on Multimedia, pp. 199–206. ACM, New York, NY, USA (1993). https://doi.org/10.1145/166266.166289
46. Fussell, S.R., Setlock, L.D., Kraut, R.E.: Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 513–520. ACM, New York, NY, USA (2003). https://doi.org/10.1145/642611.642701
47. Kinect for Xbox One|Xbox, https://www.xbox.com/en-US/accessories/kinect. Accessed 19 Sep 2017
48. Wang, I., et al.: EGGNOG: a continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 414–421 (2017). https://doi.org/10.1109/FG.2017.145
49. Watson Speech to Text, https://www.ibm.com/watson/services/speech-to-text/. Accessed 16 Sep 2017
50. Speech API – Speech Recognition, https://cloud.google.com/speech/. Accessed 18 Sep 2017

51. Goldin-Meadow, S.: The two faces of gesture: language and thought. Gesture **5**, 241–257 (2005). https://doi.org/10.1075/gest.5.1.16gol
52. Schober, M.F.: Spatial perspective-taking in conversation. Cognition **47**, 1–24 (1993). https://doi.org/10.1016/0010-0277(93)90060-9
53. Whittaker, S.: Things to talk about when talking about things. Hum. Comput. Interact. **18**, 149–170 (2003). https://doi.org/10.1207/S15327051HCI1812_6
54. Kraut, R.E., Fussell, S.R., Siegel, J.: Visual information as a conversational resource in collaborative physical tasks. Hum. Comput. Interact. **18**, 13–49 (2003). https://doi.org/10.1207/S15327051HCI1812_2