An Evaluation Framework for Multimodal Interaction

Nikhil Krishnaswamy and James Pustejovsky
Brandeis University
Waltham, MA, USA
{nkrishna, jamesp}@brandeis.edu

Abstract

In this paper we present a framework for evaluating interactions between a human user and an embodied virtual agent that communicates using natural language, gesture, and by executing actions in a shared context created through a visual simulation interface. These interactions take place in real time and demonstrate collaboration between a human and a computer on object interaction and manipulation. Our framework leverages the semantics of language and gesture to assess the level of mutual understanding during the interaction and the ease with which the two agents communicate. We present initial results from trials involving construction tasks in a blocks world scenario and discuss extensions of the evaluation framework to more naturalistic and robust interactions.

Keywords: multimodal, evaluation, semantics, objects, events, language, gesture, simulation, action, vision

1. Introduction

As the field of computational linguistics evolves and more sophisticated natural language systems become integrated with everyday use, naive users will come to expect their interactions to approximate what they are familiar with when communicating with another human, multimodally. With increased interest in multimodal interaction comes a need to evaluate the performance of a multimodal system on all levels with which it engages the user. Such evaluation should be modality-agnostic and assess the success of communication between human and computer, based on the semantics of objects, events, and actions situated within the shared context created by the human-computer interaction.

We use the modeling language VoxML (Pustejovsky and Krishnaswamy, 2016) as the platform for modeling the aforementioned objects, events, and actions, and use the VoxML-based simulation implementation VoxSim to create the environment in which a multimodal interaction involving natural language and gesture takes place. This allows us to exercise VoxML object and event semantics to assess conditions on the success or failure of the interaction.

2. Multimodal Interaction

A wealth of prior work exists on the role of gestural information in human-computer interaction. "Put-that-there" (Bolt, 1980) included deixis for disambiguation, and inspired a community surrounding multimodal integration (Dumas et al., 2009; Kennington et al., 2013; Turk, 2014).

As speech and gesture are processed partially independently (Quek et al., 2002), using both modalities complementarily increases human working memory and decreases cognitive load (Dumas et al., 2009). Visual information has been shown to be particularly useful in establishing mutual understanding that enables further communication (Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991; Dillenbourg and Traum, 2006; Eisenstein et al., 2008a; Eisenstein et al., 2008b). We will hereafter refer to this type of shared understanding as "common ground," which can be expressed in multiple modalities.

Coordination between humans using non-verbal communication (cf. Cassell (2000), Cassell et al. (2000)) can be adapted to the HCI domain, particularly in the context of shared visual workspaces (Fussell et al., 2000; Kraut et al., 2003; Fussell et al., 2004; Gergle et al., 2004). Allowing for shared gaze has been shown to increase performance in spatial tasks in shared collaborations (Brennan et al., 2008), and the co-involvement of gaze and speech have also been studied in interaction with robots and avatars (Mehlmann et al., 2014; Skantze et al., 2014; Andrist et al., 2017).

In the context of shared physical tasks in a common workspace, shared perception creates the context for the conversation between interlocutors (Lascarides and Stone, 2006; Lascarides and Stone, 2009b; Clair et al., 2010; Matuszek et al., 2014), and it is this shared space that gives many gestures, such as pointing, their meaning (Krishnaswamy and Pustejovsky, 2016a). Dynamic computation of discourse (Asher and Lascarides, 2003) becomes more complex with multiple modalities but embodied actions (such as coverbal gestures) fortunately do not seem to violate coherence relations (Lascarides and Stone, 2009a).

Prior work on multimodal evaluation also includes evaluation of gestural usage, although in this case *gesture* often refers to interfaces with multimodal displays, such as those on mobile devices (Oviatt, 2003; Lemmelä et al., 2008; Johnston, 2009). Evaluation of embodied virtual agents is often focused on the agent's "personality" or non-verbal actions, to help overcome the "uncanny valley" effect (Krämer et al., 2007). However, recent developments in multimodal technology and robotics provide resources on formally evaluating the success of multimodal grounding operations (e.g., Declerck et al. (2010), Hough and Schlangen (2016), Zarrieß and Schlangen (2017)), or of interactive systems (e.g., Fotinea et al. (2016)).

Many of the newest methods rely on datasets gathered using high-end technology and complex experimental setups, including motion capture, multiple depth cameras, range-finding sensors, or geometrically-calibrated accelerometry (systems rarely rely on all of these as that would be prohibitive). Our evaluation scheme is intended to be situation-agnostic and relies solely on logging the time and nature of interactions between interlocutors, conditioning on semantic elements during post-processing. In addition, using a suite of gesture-recognition software running on Titan X/Xp GPUs, the experimental setup we use relies only on

a single depth camera, a tablet computer and any machine capable of running the Unity-based VoxSim virtual world. With access to the GPUs over the internet, all required components can be minimally carried in a backpack and deployed anywhere with a fast internet connection. Coupled with a streamlined evaluation scheme, this allows data to be collected in a variety of different situations and conditions. We are concerned with evaluating a system for its effectiveness in creating mutual understanding between the human and virtual agents. An effective multimodal system should therefore support multimodal commands and shared perception, and approximate peer-to-peer conversations. We propose a semantically-informed evaluation scheme and a sample scenario for evaluation, with the expectation that a lightweight scheme for evaluating lightweight systems should scale to domain-agnostic interactions.

2.1. Gestures

Visual gesture recognition has long been a challenge for real-time systems (Jaimes and Sebe, 2007; Rieser and Poesio, 2009; Gebre et al., 2012; Madeo et al., 2016). In our demonstration system, we use Microsoft Kinect depth sensing (Zhang, 2012) and ResNet-style deep convolutional neural networks (DCNNs) (He et al., 2016) implemented in TensorFlow (Abadi et al., 2016). The system is capable of recognizing 35 independent gestures, chosen for their frequent occurrence in a prior elicitation study on human subjects (Wang et al., 2017a; Wang et al., 2017b). Seven of these are currently used in the sample blocks world task:

1. Engage. Begins the task when the human approaches the avatar, and ends it when they step back.
2. Positive acknowledge. A head nod or a "thumbs up." Used to signal agreement with avatar's choice or answer a question affirmatively.
3. Negative acknowledge. A head shake, "thumbs down," or palm-forward "stop" sign. Signals disagreement with a choice or negative response to a question.
4. Point. Deixis includes the direction of the hand and/or arm motion: one or two of front, back, left, right, up, or down. Indicates a region or object(s) in that region.
5. Grab. A "claw," mimicking grabbing an object. Tells the avatar to grasp an indicated object.
6. Carry. Moving the arm in a direction while the hand is in the grab position. "Carry up" can be thought of as pick up, while "carry down" is equivalent to put down.
7. Push. A flat hand moving in the direction of the open palm. Like "carry," but without the up and down directions. A beckoning motion signals the avatar to push an object toward the human.

Each gesture is assigned a compositional, underspecified semantics. We treat gestures as a special case of the VoxML entity type PROGRAM. Figure 1 shows an example.

Each gesture is linked to a VoxML verbal PROGRAM (e.g., the gesture in Figure 1 would also link to the verb [[PUSH]]). Each gesture and associated programs are distinguishable based on minimal pairs of features (e.g., [[PUSH]] in this vocabulary requires that fingers be pointed forward whereas [[CARRY]] keeps the fingers curved). This allows an evaluation scheme to correlate specific successes

$$\begin{bmatrix} \textbf{push} & \\ \text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{push} \\ \text{TYPE} = \textbf{gesture} \\ \text{SEM} = \textbf{push} \end{bmatrix} & \\ \text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{process} & \\ \text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:hand} \\ \text{A}_3 = \textbf{z:finger+} \\ \text{A}_4 = \textbf{v:vector} \end{bmatrix} & \\ \text{BODY} = \begin{bmatrix} \text{E}_1 = [while([open(y) \wedge fwd(z)], \\ \quad move(x, y, align(v))] \end{bmatrix} & \end{bmatrix} & \end{bmatrix}$$

Figure 1: Sample gesture voxeme: [[PUSH]]

or failures within an interaction with the features surrounding a gesture that occurred at the same time. An example enabled by evaluation might be discovering that straight-fingered gestures are more easily interpretable than curved-fingered gestures, or that gestures without direction are less ambiguous than gestures with it.

2.2. VoxSim

The virtual avatar and blocks world are implemented with VoxSim, a semantically-informed reasoning system (Krishnaswamy and Pustejovsky, 2016b) that allows the avatar to react to gestural events with both actions and words. VoxSim is built on the platform created by VoxML "voxemes," or semantic visual objects, and therefore allows the direct interpretation of gestures mapped through dynamic semantics. A sample voxeme for a gesture is given above in Section 2.1. The system also accepts speech input for simple directions, answers to yes/no questions, and object disambiguation by attributive adjective. Further information about voxeme properties is laid out in Pustejovsky and Krishnaswamy (2016).

2.3. Scenario

The sample interaction is adopted from functionality presented in Krishnaswamy et al. (2017). In this scenario, a human and an avatar in the VoxSim environment must collaborate to complete a simple construction task using virtual blocks that are manipulated by the avatar. The human has a plan or goal configuration that they must instruct the avatar to reach using a combination of gestures and natural language instructions. The avatar in turn communicates through gestures and natural language output to request clarification of ambiguous instructions or present its interpretation of the human's commands. The human may indicate (point to) blocks and instruct the avatar to slide and move them relative to other blocks or relative to regions of the virtual table. The human must also respond to the avatar's questions, when the avatar perceives an ambiguity in the human's instructions.

3. Hallmarks of Communication

As our goal in developing multimodal interactions is to achieve naturalistic communication, we must first examine what we mean by and desire out of an interaction such as that illustrated in Section 2.3.

We take the view that a "meaningful" interaction with a computer system should model certain aspects of a similar interaction between two humans. Namely, it is one where each interlocutor has something "interesting" to say, and one that enables them to work together to achieve common

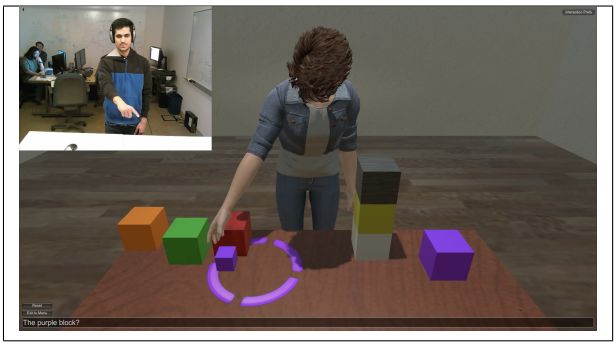

Figure 2: Example interaction setup, showing human superimposed in upper left on avatar and VoxSim

goals and build off each other's contributions, thereby conveying the impression to the user that the computer system is experiencing the same events. We therefore build the evaluation scheme off of the following qualitative metrics:

1. Interaction has mechanisms to move the conversation forward (Asher and Gillies, 2003; Johnston, 2009)
2. System makes appropriate use of multiple modalities (Arbib and Rizzolatti, 1996; Arbib, 2008)
3. Each interlocutor can steer the course of the interaction (Hobbs and Evans, 1980)
4. Both parties can clearly reference items in the interaction based on their respective frames of reference (Ligozat, 1993; Zimmermann and Freksa, 1996; Wooldridge and Lomuscio, 1999)
5. Both parties can demonstrate knowledge of the changing situation (Ziemke and Sharkey, 2001)

Many of these metrics are subjective, so we approach evaluation from a semantics-centered perspective, and use distinct semantic properties of specific elements in the interaction to determine what about the interaction enabled or hindered "shared understanding."

4. Evaluation

A robust multimodal evaluation scheme should be able to be applied to a human-computer interaction on a novel system and return a result representative of the system's *coverage* of the total possible interactions within the system's domain (e.g., construction tasks in a blocks world).
To this end, a user must be truly naive, having very little to no knowledge of exactly what the system understands. This way, during evaluation, the user has their own definition of objects, actions, and events, and the system has its own. By interacting with the system, the human should be able to learn its ways of acting within the task domain, what it knows, and what it does not.

4.1. Logging

The interaction consists of "moves" taken by each participant, which are logged live during the interaction. For instance, in our sample scenario, we log:

- Gesture received by computer (i.e., made by human)
- Word received by computer (i.e., spoken by human)
- Gesture made by computer
- Action taken by computer
- Utterance made by computer

Each of these have VoxML semantics assigned to their content.
An interlocutor's understanding throughout the task can be inferred from the logged data by the time and number of moves taken to successfully communicate an instruction (e.g., successfully indicating a distinct object, successfully indicating an action to be taken, or any combination of the previous) or understanding of an instruction (e.g, acknowledging receipt of an instruction, asking a clarifying question, or executing an interpreted action). Longer time between steps indicates more time needed for the human to think (either for interpretation or planning), or for the avatar to generate an interpretation of the human's input, and more moves in the course of completing a single instruction indicates difficulty in communicating intended meaning.
Having identified some proxy measures for the respective understanding of the human and the computer in the interaction, the task then becomes quantifying how much the two accord.

4.2. Data collection

We ran preliminary trials on the gesture interaction system heretofore described, with 20 individual human subjects, all computer science graduate students without prior knowledge of the avatar's conceptual vocabulary. All subjects were presented with a scene with six uniquely-colored, identically-sized blocks and told to collaborate with the avatar to build a six-block, three-stepped staircase. They were told the computer was capable of understanding gestures and speech but were not given specific words or gestures to use. We then logged each interaction from start to finish, defined as the point at which the human and computer had created a three-stepped staircase to the satisfaction of the human user.

4.3. Post-processing

Consider the following output snippet from a sample trial. Beginning with the block setup in Figure 2, the human attempts to instruct the computer to select the purple block and slide it next to the yellow block. Moves are numbered, `G` indicates gestures, `S` indicates spoken words, `A` indicates actions taken, and `P` indicates a continuously-updated moving coordinate interpreted from the human user's pointing and situated in the virtual world, as indicated by the purple circle in Figure 2. Each update in continuous pointing is interpreted as a separate move and multiple updates are elided in the example below for space reasons. Prefixed `H` indicates a human move while prefixed `A` indicates an avatar move. The description and interpretation of each gesture is laid out in Krishnaswamy et al. (2017) and above in Section 2.1. Time format is given as seconds after trial start.

```
1 HG engage start 1.145281
2 AS "Hello." 1.145281
3 HP r,-0.25,-0.87 4.889832
...
73 HP r,-0.08,11.69 8.552608
74 HG right point high,-0.02,5.45
  8.588802
75 AS "Are you pointing here?" 8.588802
```

```
76 AG look_at(<0.03367592; 0.743948;
  1.559844>) 8.588802
77 AG point(rHand,<0.03367592; 0.743948;
  1.559844>) 8.588802
78 HP r,-0.02,5.45 8.660162
...
126 HP r,-0.14,-0.62 11.97283
127 HS NO 12.03008
128 AS "Sorry, I don't know what you
  mean." 12.03008
129 HP r,-0.10,-0.41 12.07262
...
188 HP r,0.96,-17.45 16.93849
189 HS PURPLE 16.97243
190 AS "OK, go on." 16.97243
191 AG reach(block7) 16.97243
192 HP r,1.06,-19.16 17.06419
193 HP r,-0.25,-0.39 17.32132
194 HP l,-0.20,1.66 18.01301
195 HP l,-0.20,1.66 18.13922
196 HP l,-0.20,1.66 18.20552
197 HG push left high 19.28887
198 HG push left stop 20.07251
199 AS "Do you want me to push this
  this way?" 20.07251
200 AG LARM_PUSH_RIGHT 20.07251
201 HG posack high 22.76486
202 AS "Yes?" 22.76486
203 AG RARM_THUMBS_UP 22.76486
204 AG HEAD_NOD 22.76486
205 HS YES 25.16582
206 AS "Should I push the purple block
  right of the yellow block?" 25.16582
207 HS YES 26.53876
208 AS "OK." 26.53876
209 AA slide(block7,left(block1))
  26.53876
```

Table 2: Sample trial log

As a proxy for the human's understanding of an avatar move, we simply take the time elapsed between the *first* in a block of avatar moves uninterrupted by a human move, and the human response that follows. The avatar's move can be gesture, action, utterance, or ensemble of multiple types, and multiple moves in a block often happen simultaneously (e.g., moves 199 and 200 here). Here, we can see that between moves 190 and 192, *0.092* seconds elapsed, whereas between 202 and 205, *2.401* seconds elapsed. These differences should reflect the clarity or expressiveness of the avatar's move. We call this the human's time to *begin response*. The avatar's utterance at 190 simply prompts the human to move to their next planned step, so the human can proceed without having to spend time interpreting it. The utterance at 202 is a question asking the human to confirm a move, which requires the human to process the preceding discourse and infer some of the computer's intent in order to respond properly, possibly accounting for the longer response time.

Often, the human may make (or the gesture recognition may see) gestures that, in the current context, the avatar has no interpretation for, and thus the human makes multiple moves before the avatar responds. These circumstances are also captured by measuring the time between the first in an uninterrupted block of moves by the human, and the first response by the avatar thereafter. Between steps 3 and 75, the human points around for *3.699* seconds before landing on a particular spot that the avatar asks to confirm. Later, *3.008* seconds pass between the human's moves beginning at 192 (responding to the avatar's utterance at 190) and the avatar's response to the subsequent content at 199, a length which may indicate difficulty moving the conversation forward. The system may be misinterpreting the gestures received or the human may be making gestures the system does not recognize. By contrast, when the human succeeds in producing contentful gestures or speech interpretable in context, the avatar is able to respond immediately as the input is processed (cf., moves 1-2, 201-202, 205-206) and move the conversation forward. We call this the avatar's time to *recognize content*. The human may have trouble communicating something contentful at the beginning, but by the time context is established through deixis and object disambiguation, the avatar is able to advance the interaction by providing a possible interpretation of the human's *push* instruction. These distinctions, if consistent across multiple trials, show areas where the communication between the interlocutors flows quickly or more slowly.

Response times may be charted against the semantic features of the moves that prompted the relevant response. As VoxML structures are componential, the distribution of response times can be plotted as a probability density over the magnitude function of preceding moves that contain a given semantic feature. For instance, the response time to a *push* instruction, where the fingers must be pointed forward (as in Figure 1), can be compared to response times to a *carry* instruction where the fingers must be curved. Response times can be divided with quantiles with a q selected for the desired granularity, and comparable moves should be those that occur in similar semantic contexts, that is, $[m_{j-n}..m_{j+n}]$ where m_j is the move in question, examined in a window of size $2n + 1$. Thus $\mathbf{P}(t_i|m_{j-n}..m_j..m_{j+n})$ represents the probability that a response time t falls in an interval i given a move and surrounding context. Individual moves can be replaced by a VoxML semantic feature of the move. Higher P for lower i indicates a higher likelihood of understanding being shared through the move or semantic feature at m_j.

5. Preliminary Results

Here we present a selection of the some of the most interesting and illustrative results drawn from the pilot user studies. In each chart, the X-axis shows the quantile in which reaction times fall relative to all reaction times per agent for any move, and the Y-axis shows the probability that the reaction time to the move in question falls within that quantile (on a 0%-100% scale). q = 5 in these plots.

Figure 3 shows the distribution of times taken for the avatar to recognize the verbal and gestural realizations of positive acknowledgment and negative acknowledgement, respectively. The distributions within modalities track each other roughly, but the avatar tends to take more time to recognize spoken "yes" than "no," which may be because the human takes more time to communicate a spoken positive acknowledgment than a negative one. Meanwhile, the

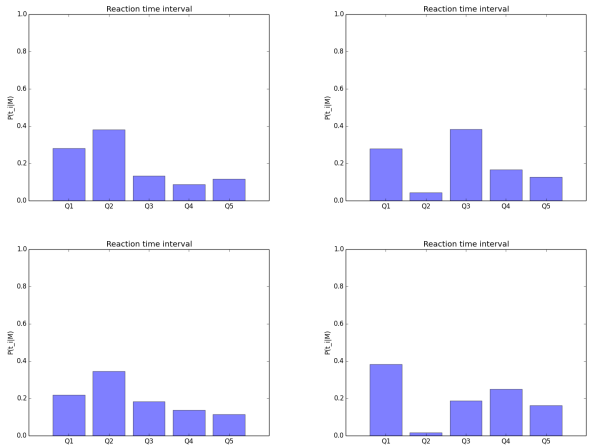

Figure 3: $\mathbf{P}(t_i|M)$, for avatar time to recognize positive/negative acknowledgment gesture (left top/bottom) vs. word "yes"/"no" (right top/bottom)

avatar appears to have a slightly quicker reaction time, in most cases, to positive acknowledgment through gesture than negative acknowledgment.

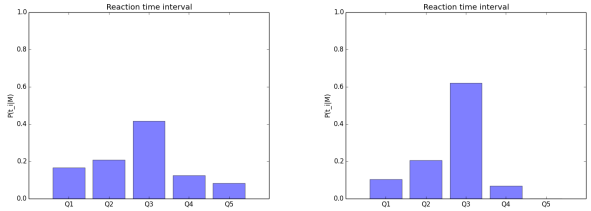

Figure 4: $\mathbf{P}(t_i|M)$ for human time to begin response to [[PUSH]] gesture (left) vs. [[CARRY]] gesture (right)

$$\left[\begin{array}{l}\textbf{push}\\ \text{TYPE} = \left[\begin{array}{ll}\text{ARGS} = & \left[\begin{array}{ll}\ldots & \\ \text{A}_2 = & \textbf{y:hand}\\ \text{A}_3 = & \textbf{z:finger+}\\ \ldots & \end{array}\right]\\ \text{BODY} = & \left[\begin{array}{ll}\text{E}_1 = & [while([open(y)\wedge\\ & fwd(z)], \ldots]\end{array}\right]\end{array}\right]\end{array}\right]$$

$$\left[\begin{array}{l}\textbf{carry}\\ \text{TYPE} = \left[\begin{array}{ll}\text{ARGS} = & \left[\begin{array}{ll}\ldots & \\ \text{A}_2 = & \textbf{y:hand}\\ \text{A}_3 = & \textbf{z:finger+}\\ \ldots & \end{array}\right]\\ \text{BODY} = & \left[\begin{array}{ll}\text{E}_1 = & [while([curved(y)\wedge\\ & down(z)], \ldots]\end{array}\right]\end{array}\right]\end{array}\right]$$

Figure 5: Abbreviated gesture voxeme type structures: [[PUSH]] vs. [[CARRY]]

Figure 4 shows the distribution of times taken for the human to respond to the avatar making a [[PUSH]] or [[CARRY]] gesture. The distributions are roughly equivalent, favoring the mid-range, but [[PUSH]] is almost twice as likely as [[CARRY]] to have a "very quick" (first interval) response time and middle-quintile response times are accordingly lower. The minimal distinction between [[PUSH]] and [[CARRY]] is the orientation and curvature of the hand and fingers, as shown in Figure 5, and so we can surmise that gestures with curved fingers might be harder for the human to interpret when compared to gestures with straight fingers (this evidences a conjecture made in Section 2.1.).

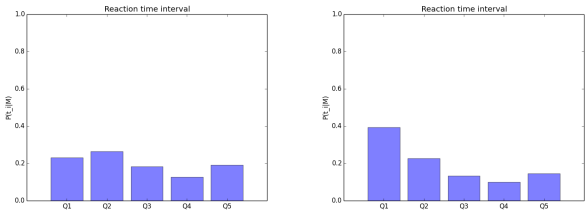

Figure 6: $\mathbf{P}(t_i|M)$ for avatar time to recognize [[POINT]] gesture with the left hand (left) vs. right hand (right)

Figure 6 shows the distribution of times taken for the avatar to recognize a pointing gesture with the left and right hand, respectively. The distributions roughly track each other, but the avatar is notably quicker to recognize right-hand pointing than left-hand pointing. Semantically, these gestures are the same, with only the orientation of the extended finger relative to the hand flipped. A possible explanation is that the gesture recognition displays greater variance in detecting the coordinates denoted by left-hand pointing, which should be accounted for in the recognition model.

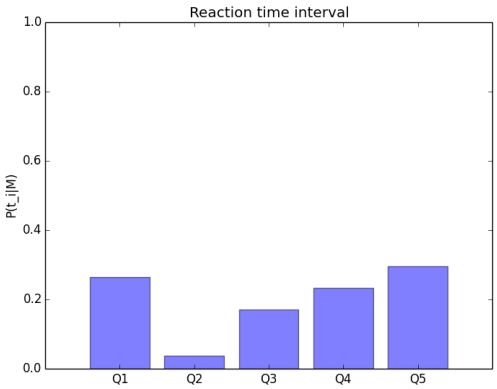

Figure 7: $\mathbf{P}(t_i|M)$ for human time to begin response to avatar asking a question

Figure 7 shows the distribution of times taken for the human to respond to the avatar's question. In most cases, the human answers quickly, or takes a long time to answer. This may be caused by the human not hearing or realizing that a question has been asked, and since the system does not repeat itself, some time can pass before the human realizes the move to make to move the conversation forward.
As developers, inferences like the above provide useful information for improving our example system.

6. Evaluation Variants

The evaluation scheme presented is deliberately high-level, in order to infer largely qualitative information from quantitative metrics. It represents a kind of minimal "base case" of conditions and parameters for evaluation of a multimodal human-computer system, many of which are easily varied to test different aspects of the interaction. As such, it is purposefully designed to be extensible to allow for different system types, interactive modalities, and scenarios.
Inference based on probabilistic distributions of course risks bias in the inferencing or due to missing information. Longer response times may occasionally be due to the trial subject getting distracted or some lag in the system rather

than any loss of understanding. Therefore we would like to present a set of variant and optional metrics to account for some of these potential omissions and test different aspects of the system under evaluation.

6.1. Additional and Optional Metrics

One simple additional parameter to log is flagging those utterances that clearly indicate confusion, such as the avatar saying "I don't understand," or the human saying "never mind," or abruptly disengaging mid-interaction. This can easily be evaluated to determine the probability of obvious confusion given a semantic context.

Similarly, affirmative and negative responses can be incorporated as a heuristic in determining when a block of moves should be initiated and terminated in evaluation. In Table 2, for instance, at move 75 the avatar asks the user if they are pointing at a particular location, and that question is answered with "No." That entire span in the interaction indicates communication of the pointing concept from the human to the avatar, but misinterpretation of the intent. If we were to treat that entire span as a block when calculating reaction time, rather than the raw human vs. avatar moves, we could glean additional information about the accuracy of pointing in the scenario.

6.2. Variant Conditions

In the outline above, we begin trials having given the user little to no information about the computer's vocabulary or capability. This allows evaluation to test the coverage of the system, but not necessarily how well the human adapts their behavior to the system. Therefore trials might also be conducted after the user has been presented with written descriptions of possible gestures (as in Section 2.1.), or after watching a video recorded from a successfully completed interaction, which would allow them to see gestures and known vocabulary in use. Both these conditions could allow evaluation of how well the interaction functions within the known constraints on the vocabulary, and can test if training the user reduces error.

6.3. User Feedback and "Ground Truth"

User reaction to the trial may be assessed by surveying them immediately following the interaction. For instance, they might rate certain moves by the avatar from clear to confusing, giving a qualitative rating of specific circumstances. Additionally, users might provide live feedback during the trial by a "talkback mode," wherein they could provide inputs such as the following:

- What user expects avatar to do following their move
- Whether an avatar question is reasonable or not
- Whether an avatar response is situationally inappropriate, incomplete, redundant, etc.

How can the overall success or failure of an interaction be assessed from the perspective of the computer system? At the beginning of the interaction, the target pattern might be fed into a planner that determines an optimal solution of moves to make on the part of both parties in order to build the target pattern (i.e., a ground truth solution), and at the completion of the interaction, the actually executed interaction is compared to the optimal solution. This can be assessed using simple metrics, like edit distance.

6.4. Scenario Variants and User Modeling

Blocks world tasks can serve as a proxy for situations requiring a collaborative interaction in a controlled environment, but do risk missing information about what a user knows about other types of complex objects versus what a computer knows. Introducing non-block objects adds other parameters that can be conditioned against, such as how the interlocutors interpret each others' behavior with convex objects vs. non-convex objects, or round objects vs. flat ones. Would a human, for example, more readily ask an agent to roll a ball than roll a block, due to knowledge that balls *afford* being rolled (Gibson, 1977; Gibson, 1979)? This information can then be incorporated into the agent's model of what its interlocutor knows about the vocabulary of available concepts. As the virtual agent becomes more and more certain that the human knows certain concepts or prefers certain moves, it may more readily execute them, or could even plan for expected user behavior.

7. Conclusion

We have proposed an evaluation scheme to assess the coverage of multimodal interaction systems and provided an outline of its use evaluating a sample interaction in a system that uses linguistic, gestural, and visual modalities. The example system exploits many advantages of virtual embodiment (Kiela et al., 2016), but consistent evaluation is required to test areas where the system needs improvement, and the framework outlined above can provide this information without very complicated algorithms to process the logged data. It uses simple metrics and processing based on object and event semantics. These properties are agnostic to the precise modalities used in a given interaction, and so the evaluation scheme accommodates measurement of various phenomena through the course of a human-computer interaction in a multimodal system. We have presented preliminary results from naive users run through the sample system, which show how we can use simple metrics to assess the ease or difficulty with which specific features communicate information. We believe this type of evaluation will be useful for developing user models and helping researchers assess the gaps in novel computational interaction systems in a variety of modalities, scenarios, and interaction types.

8. Acknowledgements

The authors would like to thank the reviewers for their helpful comments. We would also like to thank Robyn Kozierok, Brad Goodman, Lynette Hirschman, and their colleagues at the MITRE Corporation for developing the "hallmarks of communication" rubric. We would like to thank our colleagues at Colorado State University and the University of Florida for developing the gesture recognition systems: Prof. Bruce Draper, Prof. Jaime Ruiz, Prof. Ross Beveridge, Pradyumna Narayana, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Jason Yu, and Jesse Smith; and our Brandeis University colleagues, Tuan Do and Kyeongmin Rim, for their work on VoxSim. This work is supported by a contract with the US Defense Advanced

Research Projects Agency (DARPA), Contract W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

9. Bibliographical References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*.

Andrist, S., Gleicher, M., and Mutlu, B. (2017). Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2571–2582, New York, NY, USA. ACM.

Arbib, M. and Rizzolatti, G. (1996). Neural expectations: A possible evolutionary path from manual skills to language. *Communication and Cognition*, 29:393–424.

Arbib, M. A. (2008). From grasp to language: embodied concepts and the challenge of abstraction. *Journal of Physiology-Paris*, 102(1):4–20.

Asher, N. and Gillies, A. (2003). Common ground, corrections, and coordination. *Argumentation*, 17(4):481–512.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Bolt, R. A. (1980). *"Put-that-there": Voice and gesture at the graphics interface*, volume 14. ACM.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477, March.

Cassell, J., Stone, M., and Yan, H. (2000). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 171–178. Association for Computational Linguistics.

Cassell, J. (2000). *Embodied conversational agents*. MIT press.

Clair, A. S., Mead, R., Matarić, M. J., et al. (2010). Monitoring and guiding user attention and intention in human-robot interaction. In *ICRA-ICAIR Workshop, Anchorage, AK, USA*, volume 1025.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Lauren Resnick, et al., editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39, February.

Thierry Declerck, et al., editors. (2010). *Semantic Multimedia: 5th International Conference on Semantic and Digital Media Technologies, SAMT 2010. Saarbrücken, Germany, December 1-3, 2010. Revised Selected Papers*.

Dillenbourg, P. and Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151.

Dumas, B., Lalanne, D., and Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction*, pages 3–26.

Eisenstein, J., Barzilay, R., and Davis, R. (2008a). Discourse topic and gestural form. In *AAAI*, pages 836–841.

Eisenstein, J., Barzilay, R., and Davis, R. (2008b). Gesture salience as a hidden variable for coreference resolution and keyframe extraction. *Journal of Artificial Intelligence Research*, 31:353–398.

Fotinea, S.-E., Efthimiou, E., Koutsombogera, M., Dimou, A.-L., Goulas, T., and Vasilaki, K. (2016). Multimodal resources for human-robot communication modelling. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Fussell, S. R., Kraut, R. E., and Siegel, J. (2000). Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 21–30, New York, NY, USA. ACM.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., and Kramer, A. D. I. (2004). Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Hum.-Comput. Interact.*, 19(3):273–309, September.

Gebre, B. G., Wittenburg, P., and Lenkiewicz, P. (2012). Towards automatic gesture stroke detection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Gergle, D., Kraut, R. E., and Fussell, S. R. (2004). Action As Language in a Shared Visual Space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 487–496, New York, NY, USA. ACM.

Gibson, J. J. (1977). The theory of affordances. *Perceiving, Acting, and Knowing: Toward an ecological psychology*, pages 67–82.

Gibson, J. J. (1979). *The Ecology Approach to Visual Perception: Classic Edition*. Psychology Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hobbs, J. R. and Evans, D. A. (1980). Conversation as planned behavior. *Cognitive Science*, 4(4):349–377.

Hough, J. and Schlangen, D. (2016). Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies.

Jaimes, A. and Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134.

Johnston, M. (2009). Building multimodal applications with EMMA. In *Proceedings of the 2009 interna-*

tional conference on Multimodal interfaces, pages 47–54. ACM.

Kennington, C., Kousidis, S., and Schlangen, D. (2013). Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. *Proceedings of SIGdial 2013*.

Kiela, D., Bulat, L., Vero, A. L., and Clark, S. (2016). Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432*.

Krämer, N. C., Simons, N., and Kopp, S. (2007). The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In *International Workshop on Intelligent Virtual Agents*, pages 238–251. Springer.

Kraut, R. E., Fussell, S. R., and Siegel, J. (2003). Visual Information As a Conversational Resource in Collaborative Physical Tasks. *Hum.-Comput. Interact.*, 18(1):13–49, June.

Krishnaswamy, N. and Pustejovsky, J. (2016a). Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.

Krishnaswamy, N. and Pustejovsky, J. (2016b). VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.

Krishnaswamy, N., Narayana, P., Wang, I., Rim, K., Bangar, R., Patil, D., Mulay, G., Ruiz, J., Beveridge, R., Draper, B., and Pustejovsky, J. (2017). Communicating and acting: Understanding gesture in simulation semantics. In *12th International Workshop on Computational Semantics*.

Lascarides, A. and Stone, M. (2006). Formal semantics for iconic gesture. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, pages 64–71.

Lascarides, A. and Stone, M. (2009a). Discourse coherence and gesture interpretation. *Gesture*, 9(2):147–180.

Lascarides, A. and Stone, M. (2009b). A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004.

Lemmelä, S., Vetek, A., Mäkelä, K., and Trendafilov, D. (2008). Designing and evaluating multimodal interaction for mobile contexts. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 265–272. ACM.

Ligozat, G. F. (1993). Qualitative triangulation for spatial reasoning. In *European Conference on Spatial Information Theory*, pages 54–68. Springer.

Madeo, R. C. B., Peres, S. M., and de Moraes Lima, C. A. (2016). Gesture phase segmentation using support vector machines. *Expert Systems with Applications*, 56:100–115.

Matuszek, C., Bo, L., Zettlemoyer, L., and Fox, D. (2014). Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pages 2556–2563.

Mehlmann, G., Häring, M., Janowski, K., Baur, T., Gebhard, P., and André, E. (2014). Exploring a Model of Gaze for Grounding in Multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 247–254, New York, NY, USA. ACM.

Oviatt, S. (2003). User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE*, 91(9):1457–1468.

Pustejovsky, J. and Krishnaswamy, N. (2016). VoxML: A visualization modeling language. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., and Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193.

Rieser, H. and Poesio, M. (2009). Interactive gesture in dialogue: a PTT Model.

Skantze, G., Hjalmarsson, A., and Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50–66, November.

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195.

Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., and Ruiz, J. (2017a). EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*.

Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., and Ruiz, J. (2017b). Exploring the use of gesture in collaborative tasks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2990–2997, New York, NY, USA. ACM.

Wooldridge, M. and Lomuscio, A. (1999). Reasoning about visibility, perception, and knowledge. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–12. Springer.

Zarrieß, S. and Schlangen, D. (2017). Deriving continous grounded meaning representations from referentially structured multimodal contexts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 970–976.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MulitMedia*, 19:4–10.

Ziemke, T. and Sharkey, N. E. (2001). A stroll through the worlds of robots and animals: Applying jakob von uexkull's theory of meaning to adaptive robots and artificial life. *SEMIOTICA-LA HAYE THEN BERLIN-*, 134(1/4):701–746.

Zimmermann, K. and Freksa, C. (1996). Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied intelligence*, 6(1):49–58.