REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
30-11-2021	Final Report	31-Jul-2015 - 31-Aug-2021

4. TITLE AND SUBTITLE
Final Report: Communication through Gestures, Expression and Shared Perception

5a. CONTRACT NUMBER
W911NF-15-1-0459

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Colorado State University - Ft. Collins
Sponsored Programs
2002 Campus Delivery
Fort Collins, CO 80523 -2002

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
67824-CS-DRP.10

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			J. Ross Beveridge
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 970-491-5877

RPPR Final Report
as of 20-Dec-2021

Agency Code: 21XD

Proposal Number: 67824CSDRP **Agreement Number: W911NF-15-1-0459**
INVESTIGATOR(S):

Name: Francisco R. Ortega
Email: fortega@colostate.edu
Phone Number: 3053056391
Principal:

Name: J. Ross Beveridge
Email: ross.beveridge@colostate.edu
Phone Number: 9704915877
Principal: Y

Organization: **Colorado State University - Ft. Collins**
Address: Sponsored Programs, Fort Collins, CO 805232002
Country: USA
DUNS Number: 785979618 EIN: 846000545

Report Date: 30-Sep-2021 Date Received: 30-Nov-2021

Final Report for Period Beginning 31-Jul-2015 and Ending 31-Aug-2021
Title: Communication through Gestures, Expression and Shared Perception
Begin Performance Period: 31-Jul-2015 **End Performance Period:** 31-Aug-2021
Report Term: 0-Other

Submitted By: J. Ross Beveridge Email: ross.beveridge@colostate.edu
Phone: (970) 491-5877

STEM Degrees: 10 **STEM Participants:** 19

Major Goals: The major goals are described in the attached - uploaded - final report document

Accomplishments: The project accomplishments are summarized in the attached - uploaded - final report document

Training Opportunities: Nothing to Report

Results Dissemination: Here are a few highlights ways our team has disseminated our findings outside the normal channel of refereed publications. A complete list of publications directly resulting from this work appears at the end of the body of the main report document.

The EGGNOG dataset which formed the basis for our initial user studies in peer-to-peer human cooperation has been made publicly available as a curated dataset: https://www.cs.colostate.edu/~vision/eggnog/

The Diana system was highlighted as one of the key technologies when a set of institutions joined together to compete for the IBM AI XPrize. This included a set of judged interactive sessions where the Diana system was presented and put through a series of exercises.

Live interactive demonstrations of the Diana system were provided at numerous DARPA venues as well as other professional gathering including the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments 2020 (Best Demo Award) and AAAI 2020

It should be noted that the summary CwC paper written by Mitre highlighted aspects of our project including in particular the need for real-time integration of verbal and non-verbal communication in order to support task domains such as blocks world. The citation to the Mitre paper is: Kozierok Robyn, Aberdeen John, Clark Cheryl, Garay Christopher, Goodman Bradley, Korves Tonia, Hirschman Lynette, McDermott Patricia L., and Peterson Matthew W., Assessing Open-Ended Human-Computer Collaboration Systems: Applying a Hallmarks Approach, Frontiers in ArtificialIntelligence, Volume 4 2021

It is also noteworthy that both CSU and Brandeis are partners in one of the new AI Institutes being headed up by the University of Colorado at Boulder (https://www.colorado.edu/research/ai-institute/). Our participation in this project is a direct outgrowth of the work we carried out as part of CwC.

Honors and Awards: The Diana system was highlighted as one of the key technologies when a set of institutions joined together to compete for the IBM AI XPrize. This included a set of judged interactive sessions where the Diana system was presented and put through a series of exercises.

Best Demonstration Award at the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments 2020.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Bruce Draper
Person Months Worked: 8.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: PD/PI
Participant: Ross Beveridge
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: Francisco Ortega

Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Nathaniel Blanchard
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Anita Bundy
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Lisa Daunhauer
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Michael Kirby
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Nikhil Krishnaswamy
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Christopher Peterson
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Jaime Ruiz
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:

National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Shaikh Shawon Arefin Shimon
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Rahul Bangar
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Mohtadi Ben Fraj
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Matthew Dragan
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Vidya Gaddy
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Quanyi Mo
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Gururaj Mulay
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 20-Dec-2021

Participant Type: Graduate Student (research assistant)
Participant: Pradyumna Kumar Narayana Rao Gari
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Dhruva Kishor Patil
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: William Pickard
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Aditya Raikwar
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Erik Ridd
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Ben Sattleberg
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Ameni Trabelsi
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 20-Dec-2021

Participant Type: Graduate Student (research assistant)
Participant: Heting Wang
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Isaac Wang
Person Months Worked: 7.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: David White
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: David White
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: David White
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 20-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 3-Accepted
Journal: IEEE Transactions in Computer Graphics in Visualizations
Publication Identifier Type: DOI Publication Identifier: 10.1109/TVCG.2020.3023566
Volume: Issue: First Page #:
Date Submitted: 11/2/20 12:00AM Date Published: 9/21/20 12:00AM
Publication Location:
Article Title: Understanding Gesture and Speech Multimodal Interactions for Manipulation
Authors: Adam Williams, Jason Garcia, Francisco R. Ortega
Keywords: Gestures, multimodal interaction, ar
Abstract: The primary objective of this research is to understand how users manipulate virtual objects in augmented reality using multimodal interaction (gesture and speech) and unimodal interaction (gesture). Through this understanding, natural-feeling interactions can be designed for this technology. These findings are derived from an elicitation study employing Wizard of Oz design aimed at developing user-defined multimodal interaction sets for building tasks in 3D environments using optical see-through augmented reality headsets. The modalities tested are gesture and speech combined, gesture only, and speech only. The study was conducted with 24 participants. The canonical referents for translation, rotation, and scale were used along with some abstract referents (create, destroy, and select). A consensus set of gestures for interactions is provided. Findings include the types of gestures performed, the timing between co-occurring gestures and speech (130 milliseconds), perceived workload...
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 3-Accepted
Journal: ACM Proceeding on Human-Computer Interactions
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 11/2/20 12:00AM Date Published: 9/1/20 12:00PM
Publication Location:
Article Title: Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation
Authors: Adam Williams, Francisco R. Ortega
Keywords: multimodal interaction, gestures, speech, elicitation, augmented reality
Abstract: This research establishes a better understanding of the syntax choices in speech interactions and of how speech, gesture, and multimodal gesture and speech interactions are produced by users in unconstrained object manipulation environments using augmented reality. The work presents a multimodal elicitation study conducted with 24 participants. The canonical referents for translation, rotation, and scale were used along with some abstract referents (create, destroy, and select). In this study time windows for gesture and speech multimodal interactions are developed using the start and stop times of gestures and speech as well as the stoke times for gestures. While gestures commonly precede speech by 81 \textcolor{black}{ms} we find that the stroke of the gesture is commonly within 10 ms of the start of speech. Indicating that the information content of a gesture and its co-occurring speech are well aligned to each other.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: First IEEE International conference on Humanized Computing and Communication (HCC 2019)
Date Received: 30-Oct-2020 Conference Date: 26-Sep-2019 Date Published: 27-Dec-2019
Conference Location: Laguna Hills CA
Paper Title: User-Aware Shared Perception for Embodied Agents
Authors: David G. McNeely-White, Francisco R. Ortega, J. Ross Beveridge, Bruce A. Draper, Rahul Bangar, Dhru
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 20-Dec-2021

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Ross Beveridge
Signature Date: 11/30/21 4:16PM

Communication through Gestures, Expressions & Shared Perception

DARPA CwC Program Project Final Report

Final Report for Contract Number: W911NF1510459

Dates Covered: Jul, 31 2015 to Aug, 31 2021

Ross Beveridge, Francisco Ortega, Jaime Ruiz,
Nikhil Krishnaswamy, and Lisa A. Daunhauer

Introduction

The Colorado State University effort as part of the DARPA Communicating with Computers (CwC) program emphasized peer to peer communication between a person and an embodied agent both of whom were engaged in solving a concrete physically paced problem. Early in the program, Colorado State University teamed up with Brandeis University and collectively created what is described by us as the Diana System. The Diana System set new performance standards for combined verbal and nonverbal communication between people and agents. It demonstrated new capabilities for asynchronous multi-modal interaction. It fostered fundamental advances in AI representational tools in order to support peer-to-peer communication grounded in shared perception. From the standpoint of Colorado State University, the shortest simplest description of what was accomplished was the creation of an embodied agent able to conduct peer to peer communication using a combination of sight and speech; in particular, the key is tight coupling of computer vision to observe the person and seamlessly combine verbal and non-verbal communication. Keep in mind that at the start of CwC all commercially deployed agents (e.g., Siri, Alexa) were effectively blind, and that even now in 2021 agents in common use remain essentially blind.

The next section will summarize our approach and significant accomplishments. It will highlight three key contributions of the work. First, the critical decision early on to carefully study how two people communicate when solving a visual task. Second, the integration of sight, speech and most fundamentally a shared perception of a physical

task environment. For this part, it will be essential to understand we are describing the joint work of Brandeis University and Colorado State University. Finally, there is the discovery in the course of our work that multimodal communication is fundamentally different from unimodal in many ways. Singling out one specific that caused us perhaps both one of our greatest challenges, and also one of our greatest opportunities to succeed, was the nearly split-second coordination between gesturing and speech that must be immediately understood by a multimodal agent.

One way in which to better appreciate this last point is to consider what is readily described as “turn taking behavior”. First a person takes an action, then the agent responds, then the person responds, and so on back-and-forth. Think about two person game playing, or for that matter, how all common consumer agents behave today. This was the approach to communication exhibited most commonly in other CwC projects. In notable contrast to the turn-taking approach, we found that strict turn taking with an agent able to hear, see and speak, felt stilted and cumbersome to a person. Instead, it was both natural and expected, that communication between the person and the agent was continuous and fluid in how verbal and non-verbal cues moved the conversation forward and created a shared context.

Project History / Summary

We went back to look at our original proposal when crafting our final presentation to DARPA at the July 2020 CwC PI Meeting. Two of the statements there turn out to motivate the work we subsequently performed - better than often is possible with basic research. To start, we found:

> “The goal of the CwC program, as we see it, is to create a new paradigm for communication between people and computers, one in which people and computers converse as equal partners in cooperative tasks. Our goal within this program is to make sure that the new paradigm includes gestures and facial expressions, not just words.”

Now looking back over the project this is exactly what we did. Through the collaboration with Brandeis University and the creation of the Diana System we built, demonstrated, and tested with naive users, a system able animate through an embodied person - an avatar - a peer able understand speech, gesture, and facial expressions, all in the context of peer-to-peer problems solving.

The other quote we found is:

> “Our approach to designing intuitive interfaces relies on elicitation studies. The idea is to bring naïve users to the blocks world table and observe them as they try to cooperatively accomplish tasks with the computer.”

In light of what followed it is hard to underestimate the importance of the decision encapsulated in this quote. Knowing what we know now about embodied agents, human computer interaction, user centered design, etc., it was ambitious of us to establish from the outset the goal of building an agent - a peer - modeled directly upon what we observed people doing in human subject elicitation studies.

Indeed, our first year of research centered primarily on the human subject studies which ultimately gave rise to the collection of the EGGNOG dataset and gave us the blueprint for how we sought to have our agent interact with a person. Here briefly is an overview of that effort

Blocks World Redux

The CwC program included an application domain for testing peer-to-peer communication based upon an old and famous AI problem, namely blocks world. Those of us with experience in computer vision were of course intimately familiar with the task, its history, and in particular its relationship to early work on both computer vision and planning. However, what was exciting about the CwC Programs take on blocks-world was the challenging focus on peer-to-peer communication. To be clear, DARPA essentially forbade any of the performers from proposing or focusing upon vision algorithms to better see blocks or planning algorithms to work out complex sequences of action.

Instead, and to be clear, this was very much to our liking. DARPA placed the emphasis entirely upon advancing the state-of-the-art in peer-to-peer communication between an agent and a person who are jointly engaged in building structures made from blocks. Much of that communication was presumed to be spoken, but for our effort in particular, we saw a huge opportunity to bring in computer vision in order that our agent might understand nonverbal as well as verbal communication. Given that as our goal, the best way to discover how peer-to-peer communication might involve combined verbal and nonverbal communication is to watch people, and thus our first year of work collecting the EGGNOG dataset.

The Elicitation Study and the EGGNOG Dataset

An important and early decision that we made was to explicitly recognize the fact that CwC was a program about human computer interaction, and that it was critical to begin with a methodology appropriately grounded in human centered design. It may seem simple, but the importance of recognizing the role of human subjects studies in work addressing human computer interaction cannot be underestimated and it remained at the heart of what we were doing throughout our effort. Concise summary of the essence of human centered design appears in Figure 1.

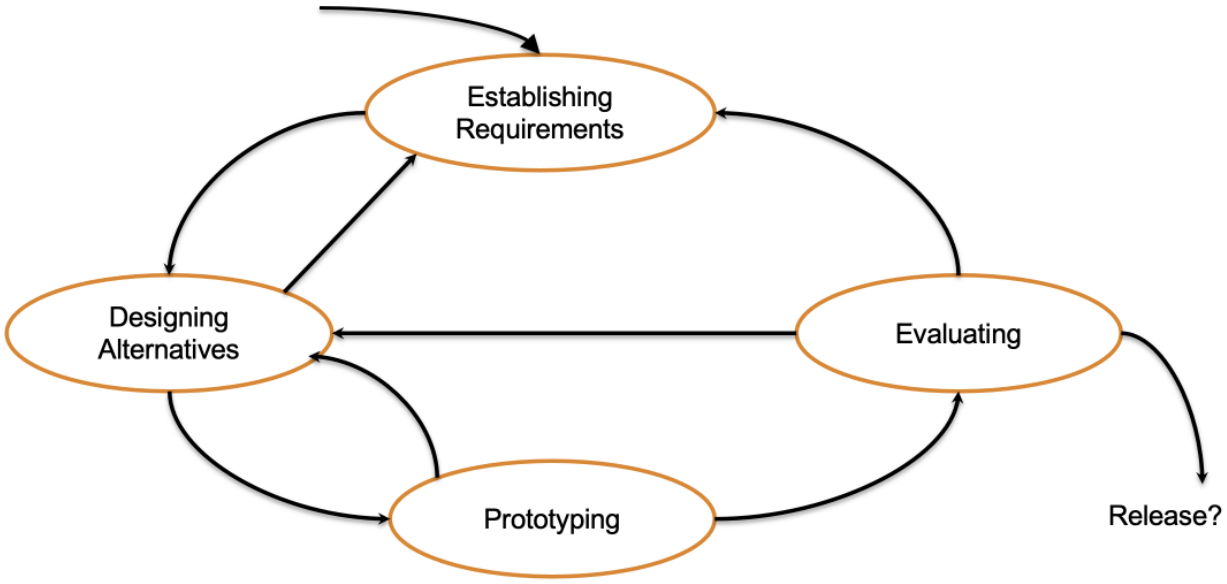
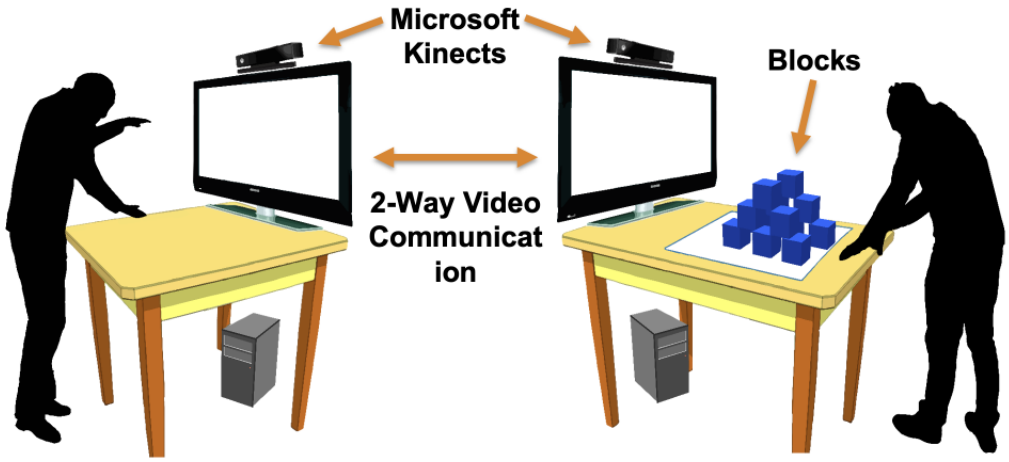


Figure 1: Human Centered Design Cycle

Looking back on the project, and informed by what our typical human subjects studies in the area of human computer interaction, it's noteworthy that we began not with a computer interface which people explored, but instead with an experiment involving two people. Recall the goal of CwC was peer to peer communication, and it followed naturally from that goal for us to design human subject experiments where a human being carried out behaviors which we sought, if we could be successful, to emulate with our agent.

Consequently, much of our first year effort was focused on producing what became known as the eggnog data set. The figure below shows the essential ideas of our set up. Two people were placed in two different rooms in front of two different tables.

Between the two rooms we had four to a video available as well as full two-way audio. Depending on the condition of the experiment we wish to run, the two people could both hear and see each other as well as seeing each other‘s table.



Conditions

Gesture Only – Audio disabled, non-verbal communication only

Speech Only – Audio is enabled, video is disabled

Speech & Gesture – Both Audio and video are enabled

Figure 2: The user elicitation design summary

One of our two subjects was designated as the signaler, and that person had a picture (off camera) of a Blocksworld structure. The other person was designated as the builder, and it was their job to actually build the structure. This experiment design forced the two people to invent ways of communicating. Experiment was carried out under three different circumstances as noted in Figure 2.

The study that we ran is broadly characterized as an elicitation study because we are eliciting responses from people in order to inform us about better system design. Being much more precise, one of our key goals was to observe the types of nonverbal communication employed by people in solving these paired person Blocks World tasks. It should also be noted upfront, that the condition in which there was no audio communication, and hence no speech, was one for which we had no certainty upfront that people would actually be able to successfully complete their task.

For our study we recruited 60 participants, mostly students, and collected over 80 hours of data. Specifically, we used a Kinect to record RGB and Depth video as well as audio of all sessions. Sessions were run in three different modes: 1) gesture only - audio was

diabled in both directions, 2) speech only - video from signaler to builder disabled, 3) speech & gesture - full bidirectional audio and video. Each participant participated in both the builder and the signaler role. Figure 3 highlights a sampling from one session.

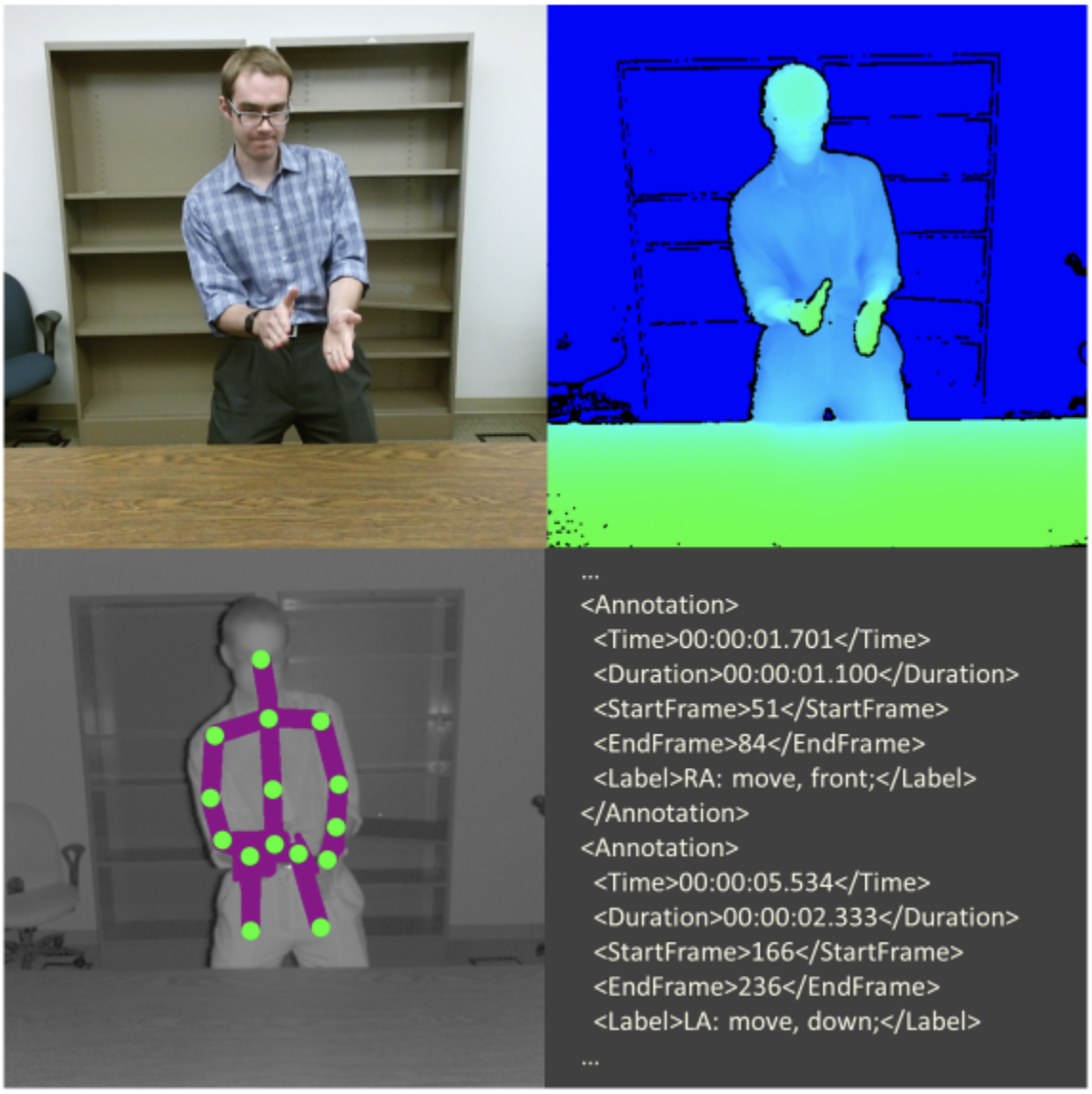


Figure 3: A sampling of data associated with the EGGNOG data collection.

It then took our lab perhaps four months with multiple graduate students and faculty members carefully annotating the rolled salting data. The final product at a low level was 24,503 distinct segmented and labeled video snippets that signified interesting movements on the part of the subjects. Movements, broadly speaking gestures, or carefully recorded for both the signaler and the builder.

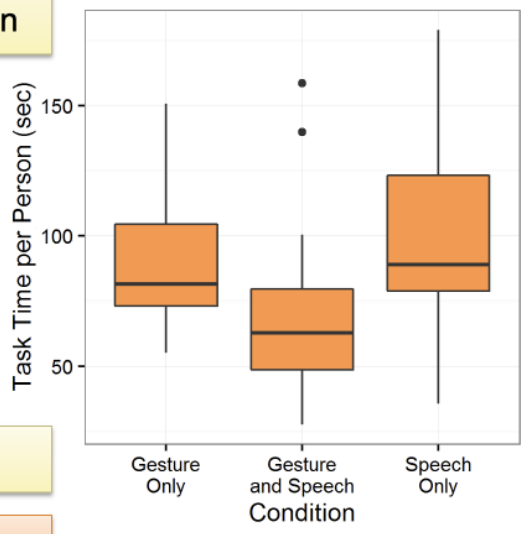
Before getting into details of what was learned from the study, it's worth highlighting that perhaps one of the most significant findings is that in all but three cases, the pairs of

people successfully built their structures. To put this another way, nonverbal communication alone is sufficient to almost always facilitate success in the types of blocks-world environments where we were seeking to place our agent.

Another key finding is summarized in Figure 4 below. What it tells us is twofold. First, and this may come as a surprise to some, there was no significant difference in task completion time for pairs of people using speech alone (no gesturing) versus people using gesturing alone (no speech). Generally there is a bit of a bias for people to assume that spoken language is clearly superior to nonverbal communication. Unquestionably, when communicating complex abstract ideas, this bias is probably fairly well-founded. However, as our data clearly shows, for very physical problem-solving it appears knowing where someone's pointing and generally what they're doing with their body is every bit as important that's what they're saying.

Looked at pairs completing block tasks
Measured average time to complete per person

Condition	Mean Time (sec)	Std. Dev.
Gesture Only	90.3	26.5
Gesture + Speech	69.6	33.5
Speech Only	97.4	39.5



Gesture only and speech only are comparable

Not surprisingly, using both together works best

Figure 4: Paired user performance summary using gesture only, speech only, and both.

The second finding is important, and is also not terribly surprising. When the pair of people tasked with building structures could both hear and see each other they solve problems a little bit faster. A little more surprising, and something which foreshadows a lot of our work in the later years, is the manner in which speech and gesture are very tightly coupled. To put this another way, often what a person is doing with their hands doesn't really make sense unless you hear what they're saying, and vice versa.

The EGGNOG dataset has been available for public download since 2017 through the following URL - https://www.cs.colostate.edu/~vision/eggnog/ (Wang et al., 2017b)

Real-Time Gesture Recognition

After careful human annotation and analysis of the EGGNOG data upwards of 40 distinct hand, arm and body poses were identified as important for peer-to-peer nonberbal communication in the blocks-world domain. Once these building blocks were identified the team turned attention toward developing novel neural networks (Narayana, Beveridge and Draper, 2018) that could recognize different hand configurations in real-time using depth imagery from the Microsoft Kinect.

The following two figures show a bit of what this system accomplishes. Figure 5 shows a person standing in front of our system being watched by the Kinect RGB and depth video. In addition, keep in mind the Kinect is creating a 3D skeleton representing the person's overall pose.

Figure 5: Example of user gesturing.

FIgure 6 is a screenshot from one of our development windows showing in real time the response for all our different basic gesture components. What you see are five panels

corresponding to right hand pose, left hand pose, right arm pose, left arm pose and finally head position. Looking down the rows of the left arm pose observe only the last label, ‘other’, has a high response. That is because the person’s right arm is down by their side and the hand pose is not one of our selected meaningful configurations. In contrast, note that the left hand response is strong for the label ‘point right’ - which is indeed what our person is doing.

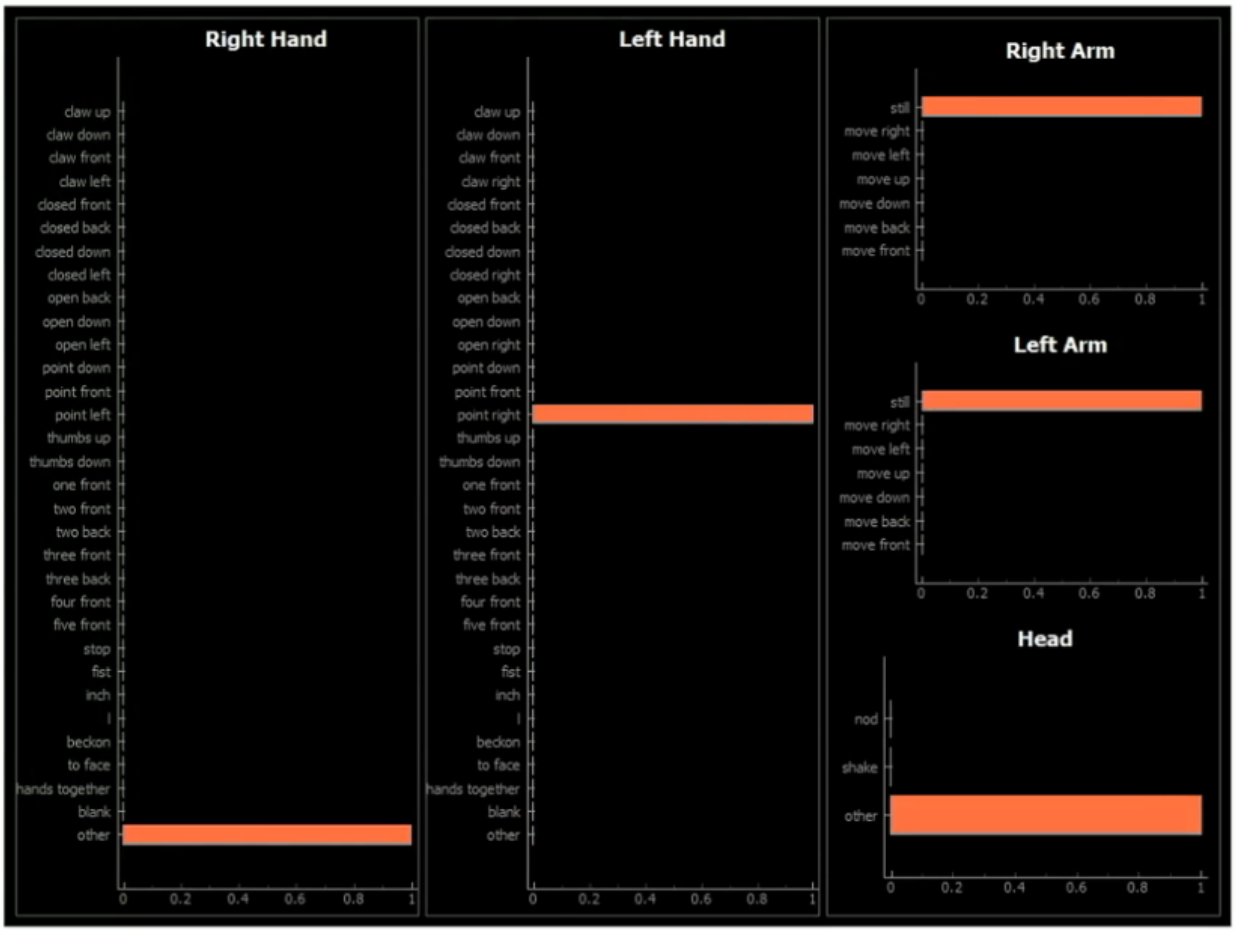


Figure 6: Real-time display of gesture recognition system

The real-time gesture recognition system is constantly running whenever our embodied agent Diana is interacting with a person.

The Diana 1.0 System

About the time CSU was assembling the computer vision components to support real-time gestural communication with our agent we also began in earnest our

partnership with the CwC team from Brandeis. Below in Figure 7 is a snapshot of that first architecture developed by CSU and Brandeis with a distinct emphasis on gesture recognition. Note that a person stands in front of a table facing a large monitor as well as the Kinect sensor. The video streams and skeleton data are then passed through a series of deep convolutional neural networks to interpret the hand and body pose of the person. To be a bit more specific, the hand poses are interpreted by the DCNNs, while body pose is interpreted directly by a combination of state-machines and 3D interpretation algorithms (not shown in the figure).

The audio stream is passed along to an off-the-shelf speech to text processor and in the earliest versions of the Diana system simple word spotting was used to identify highly context dependent communicative phrases such as “yellow block”. Finally, as noted in the bottom center of the flow, marked Semantic “Packages”, the CSU system put together semantic directives/updates which then were fed to the Brandeis system. That system, which itself is far more complex than this short summary might suggest, watches for these packages and takes appropriate actions. Keep in mind the virtual table, the Diana avatar, and the abstracted representation of the task all reside within the Brandeis system. Further, the Diana system is deeply supported by, VoxSim, a virtual platform running on the Unity gaming engine (Krishnaswamy and Pustejovsky, 2016).

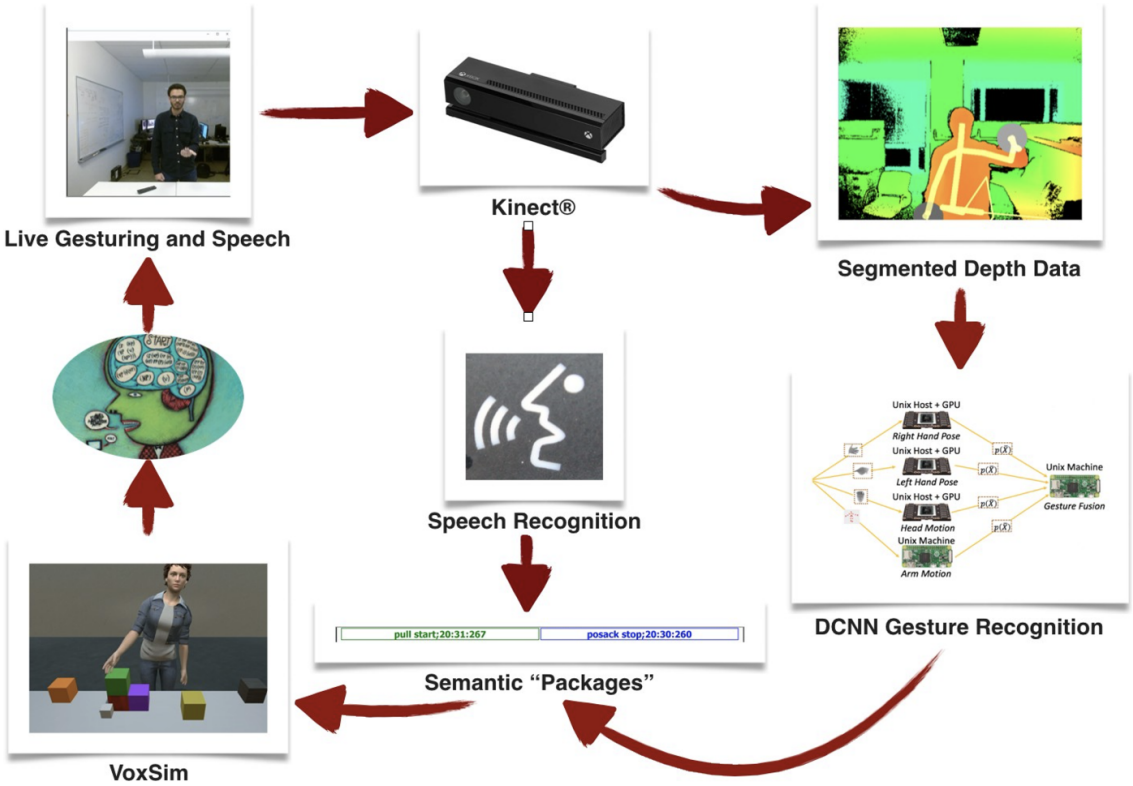


Figure 7: Snapshot of Diana 1.0 Architecture

Seeing Diana 1.0 in Action

It is very difficult to convey on printed text just how interactions with Diana 1.0 unfold. Throughout the CwC program videos were presented at meetings and shared with DARPA. Here is a video showing a session where a user interacts with Diana 1.0 to build a staircase. Figure 8 is a snapshot from the video.

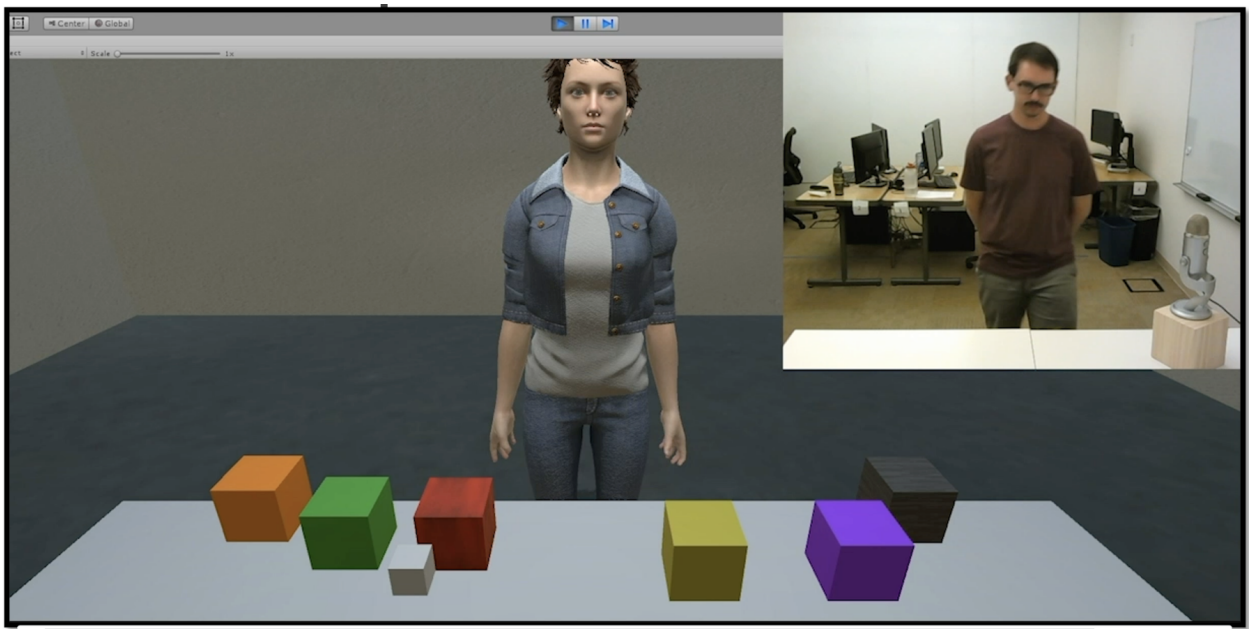


Figure 8: Snapshot from a video showing person working with Diana 1.0

The actual video can be found here:

[Link to Diana 1.0 abbreviated session]

At the highest level of abstraction notice the parallel with the EGGNOG data collection setup. Diana is standing in front of her table looking at the user, who in turn is standing in front of his table looking at Diana. Diana has all the blocks, and will be the builder, and the person takes the role of signaler. To provide just a hint at the interactions, based directly on what we observed people doing when working together, the signaller

engages the other person - or in this case Diana - by approaching the table. From that moment forward Diana watches, listens, and responds to verbal, nonverbal, and combined verbal and nonverbal communication initiated by the person.

CSU and Brandeis have created a website where you may go at this point should you wish to actually see how these interactions unfold. The site is: http://www.embodiedhci.net/. Videos linked in this report will also be hosted in perpetuity on this site beginning in the near future.

We should also point out - with some pride - that the Diana system was highlighted as one of the key technologies when a set of institutions joined together to compete for the IBM AI XPrize. This included a set of judged interactive sessions where the Diana system was presented and put through a series of exercises. Next, as we will discuss more in the following section, Diana 1.0 was put through a series of Naive User Studies. These studies were overseen by Jaime Ruiz at the University of Florida. To be clear, Jaime Ruiz was part of the CSU effort from the beginning and his participation and funding continued even after he moved from CSU to the University of Florida.

The 2018 Diana 1.0 Naive User Studies.

Studies carried out at the University of Florida in 2018 were designed to focus on two key questions.

1) Are naïve users able to interact with Diana to build specific layouts without prior instruction?

2) If Diana performs gestures while asking questions to the participant, does it result in the participants performing gestures more frequently and increase the breadth of gestures used?

For the studies a total of 45 participants were recruited. Most were students at the University of Florida. The study was run with two different variants of the Diana system. The first is what we refer to as “non-teaching“. This might be more aptly thought of as a default behavior. The other is characterized by “teaching“ which is a motive interaction that grew out of our curiosity as to how people would learn to make good use of the system. Put simply, “teaching” mode included Diana keeping track of such things as whether a person used a pointing gesture, and if not, then would use pointing herself to suggest she understood what pointing meant.

Each participant completed 8 trials for a total of 360 trials in which a participant was asked to build a structure working with Diana. Participants were only given the following information about the system:

“This is Diana. She is a virtual agent that can understand both speech, gesture, and pointing in order to manipulate blocks on the table. You must work with Diana to complete the tasks.”

Instructions provided to participants were intentionally left vague, as our goal is to capture the use of the system from naïve users. The results are summarized in Figure 9 below. The block patterns the users were asked to construct of course were varied. A trial ended either when the pattern was correctly constructed or after 10 minutes. Out of all 360 trials, participants failed in the sense that they reached their 10 minute time limit in 26 cases. Hence, 93% of the time Naive users were able to construct the specified pattern.

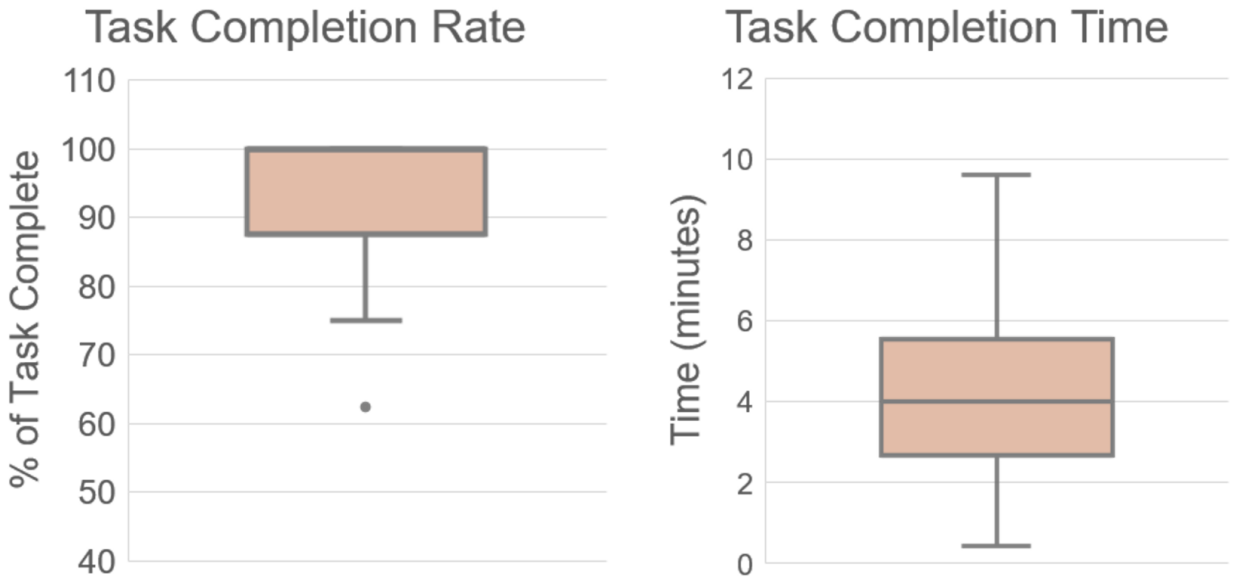


Figure 9: Task complete rates and times for 2018 naive user studies.

As the task completion rates and times clearly show, naive users are able in the vast majority of cases to work with Diana to build block structures.

The question about teaching users through example also suggests the behavior did actually help users increase their use of gestures and speech. Figure 10 below summarizes our findings.

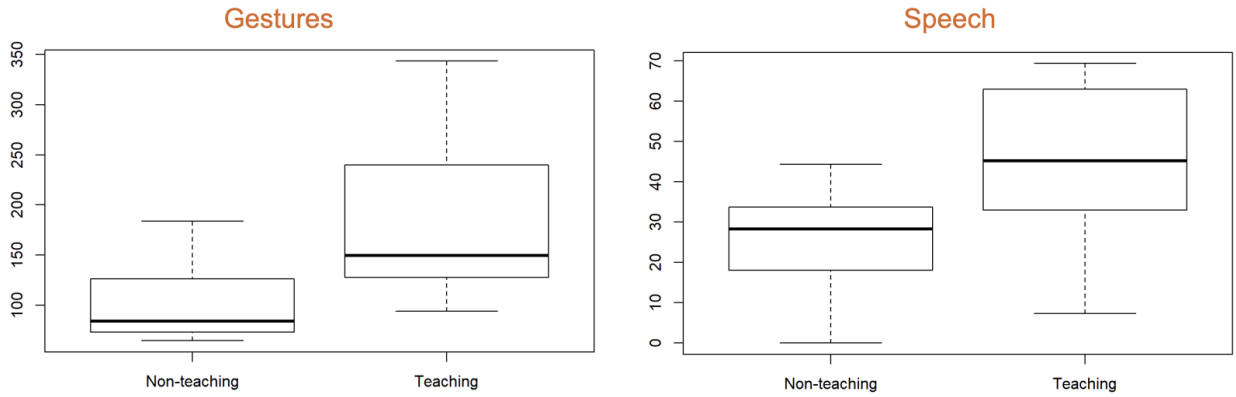


Figure 10: User uptake for gestures and speech associated with Teaching Mode.

In Figure 10 the vertical axis in both cases reflects how often a user used gestures - left plot - or speech - right plot - when working with the non-teaching versus the teaching system. These experiments are early relative to the broader questions of how an agent can subtly help a person better communicate. Nonetheless, the results clearly show people do pick up on cues from the agent and can adapt their behavior in response to how they see the agent communicating with tem.

So far the arc of our project, from concept through development of a full interactive system modeled on human-to-human communication, through to formal naive user evaluation, was extremely encouraging. Indeed, Diana 1.0 recognized gestures and integrated verbal and nonverbal communication. Ultimately, people could use the system to accomplish tasks.

However, there is one more important element to the 2018 study. As is common when following a user centered design methodology, all participants completed a System Usability Scale (SUS) questionnaire after their encounter with Diana. The mean SUS score was 41.0. To put this in perspective, a score of 68 is considered average and 80 highly useful. Put simply, our Naive users succeeded over 90% of the time, but they clearly did not fully enjoy the experience.

The question coming out of these studies is what can be done to improve the user experience, and our own hands-on practice with Diana 1.0 along with user comments gave a pretty solid indication where to pay more attention. Here are two user comments included as part of their SUS responses:

> “Sometimes the AI would loop through a long menu of actions she could perform, but none of them would be what I wanted to do, so I felt like there was some time wasted just trying to get back to a point where I could try a command again.” -P5

> “Sometimes it was difficult to communicate with Diana. She would ask a series of questions automatically rather than just listen to your command.” -P111

These comments bolstered our own observation that while careful and meticulous about asking for confirmation, Diana 1.0 was correspondingly annoying. The following is a bit of a caricature of a common interaction, but it goes like this. User says “Put the yellow block on the red block”. Diana responds: “Do you want me to put the yellow block on the red block?” It is not hard to see the problem. We as people grow quickly tired of this apparent lack of confidence and - from our standpoint - needless redundant communication. However, while at one level easy to spot as a weakness, fixing the problem turned out to be far more involved and far more suggestive of how multimodal agents must behave if they are to get along with people.

Diana 2.0 - Asynchrony, Initiative and Affect

Like most of the interactive systems developed in CwC, Diana 1.0 was for the most part a turn taking system. Generally, one used speech, gesture, or a combination of the two, to essentially issue a directive. Then, Diana would take her turn and either ask a clarifying question or carry out an action. There were a handful of exceptions to this strict turn taking. For example, Diana 1.0 understood “never mind” - which could be conveyed either through speech or gesture (a palm-out, “stop” gesture). It turns out being able to get an agent’s attention while it is proceeding to do something you wish it would not do is extremely important, and with “never mind” Diana 1.0 could at least stop what she was doing and look at the person attentively - ready to do something new and hopefully more to their liking. While good and essential, this was a very limited example of something much bigger and more important.

To illustrate, consider the following scenario. A person instructs Diana “Put the yellow block on the red block.” Diana proceeds to carry out this sequence of actions, picking up the yellow block, looking at the red block (yes, Diana uses gaze to communicate understanding extralinguistically) and starts moving the yellow block toward the red block. At that point the person says, “No, there” while pointing at a nearby green block. This is an asynchronous interruption of what Diana is focused upon doing. Yet, were Diana a person, there would be no doubt what should happen. She should simply adapt her motion to put the yellow block on the green block. Hopefully this example is easy to visualize and understand. It was not simple to implement with Diana 1.0. Indeed, supporting this level of fine grained integration of sight and speech along with split second timing took a nearly complete rebuild of the Diana system.

The Diana 2.0 Architecture

The rebuild began with a serious rethinking of how communication between components would be supported in the new architecture. What emerged was a new design that leveraged a powerful and relatively old idea from AI, namely the blackboard (Erman et al., 1980). Proposed in the 1970s for a natural language understanding system a blackboard architecture is pretty much what the name implies. Imagine a set of asynchronous processes all able to write on a common blackboard, and at any time read what is written on the blackboard. Further, imagine that chalk is used to mark off portions of the blackboard associated with communication related to specific topics, for example, gestures. This idea became the basis for much of the Diana 2.0 rebuild (Strout, 2020).

In parallel with the decision to go with a blackboard architecture came the decision to become much more reliant upon VoxSim's event management mechanisms. To put this another way, it's not possible to implement a capable blackboard architecture without the benefit of a solid event handling backbone. In terms of lessons learned from our project, this expansion of our reliance upon a mature gaming engine should not be underestimated. As has been well described by the Brandeis team elsewhere, the Unity Gaming Platform came into the Diana project because it allowed for an integration of symbolic and physical (physical simulation) representation as the basis for natural language understanding, i.e., what is often referred to as VoxSim. This use of Unity also dovetailed well with the need for a real-time 3D physical simulation of both the agent - Diana - and her world - the table and blocks. With the transition to Diana 2.0, the reliance upon the Unity Platform grew, even further, tapping into the capable and well proven event driven processing capabilities built into Unity. An update schematic of the Diana 2.0 system is shown Figure 11 below.

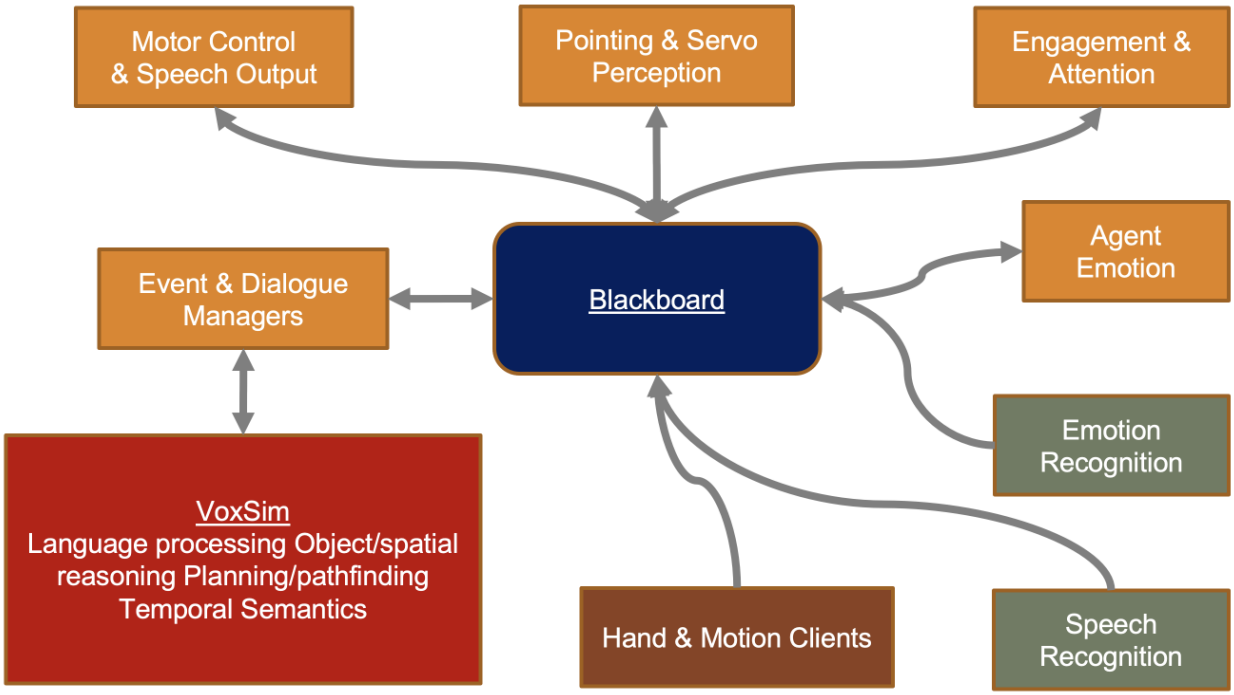


Figure 11: Diana 2.0 Architecture with blackboard at the center.

Example 1: “Wait, on the white one”

With the upgraded architecture for Diana 2.0 it became possible for the Diana system to be reactive to both speech and text in a manner that was timely and enabled asynchronous communication between a person and the agent. The change represented a significant milestone in terms of moving away from turn taking. The new capability can be seen in a video for which the link is included here. Figure 12 shows a screenshot from this video; the link to the actual video follows:

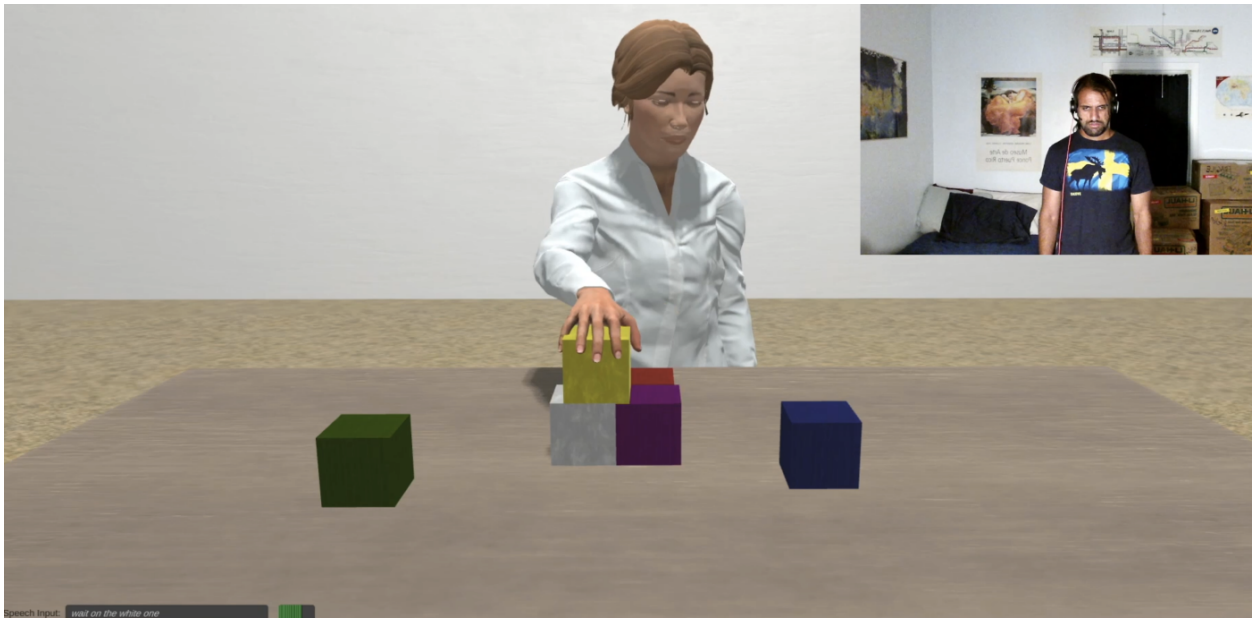


Figure 12: Example of Diana being corrected in the middle of an action

Wait, on the white one

What you will see is Diana receives the following through speech “Put the yellow block there.” While the person is speaking they are also pointing out a block on the table. Diana proceeds to move the yellow block on top of the purple block. The purple block was indeed where the user was pointing. But as she’s completing this action the person says “wait, on the white one.“ In response, Diana smoothly moves the yellow block from on top of the purple block to on top of the white block.”

Honestly, we as people take so much about multimodal communication between people for granted that at first this interaction may not seem significant. Afterall, two people could accomplish what we just described and without ever even being consciously aware of how we as people handle integration of gesture and speech, not to mention succinct requests to refine an action. However, we are not aware of any other multimodal agent system today able to carry out this level of integration between speech, gesture, and task context. The following figure is a screen shot of Diana moving the yellow block over to rest atop the white block that takes place during this interaction.

Example 2: “No, the white one.”

Another short example of how smoothly gesture and speech are integrated into actions in a manner that is asynchronous and fluid can be seen in the following video.

No, the white one

What one observes in this video is that the person begins by pointing at a block on the table. The specific block being pointed to is the yellow block and Diana responds by placing her hand over the yellow block. The agent is indeed taking initiative at this point and anticipates a coming request to do something with that block. The person then says “no, the white one“ while proceeding to point out another position indicating where to place a block. Specifically the person points at the top of a blue block. Upon hearing “no, the white one“, Diana shifts her hand position to reflect understanding of the statement. In other words, she places her hand over the white block. Being a multimodal system, Diane also offers verbal confirmation by saying “OK,” and she then proceeds in a smooth motion to move the white block onto the top of the blue block. No more speaking on the part of Diana, she simply puts her arms to her side and looks back at the user. This is an example of the ability of Diana 2.0 to simultaneously pay attention to speech, gesture, and asynchronously attend to both in the course of carrying out an action.

Recall when we discussed the system user satisfaction results from our first naïve user studies, just how much people disliked repeated questions that felt obvious, such as “do you want me to put the white block on the blue block?“. As already noted, these changes in behavior that we’ve implemented seem somewhat simple and obvious after the fact. But they’re quite significant, and in the specific case of the “no, the white one“ example, what we’re seeing is an instance of Diana following a very different strategy for peer to peer human computer interaction. In the Diana of 1.0 system, the agent was very cautious and did not want to make a mistake. That was in part because correcting a mistake involved a somewhat tedious additional set of turn-taking style exchanges between the person and the agent. What’s evident and important in the current example is that an agent should simply take the initiative with its best guess of what is desired and listen carefully for redirection. If the redirection is received, then it is immediately acted upon and consequently the overall accomplishment of what the user desires shall feel smooth and simple.

Example 3: The Multimodality Spectrum

In discussing the capabilities of the Diana 2.0 system, there's one more accomplishment that merits serious attention. Recall that in our original human to human subject studies we discovered that people could solve Blocksworld tasks using only gesture, only speech, and with a small improvement in task completion time, people would use a combination. With the changes associated with the Diana 2.0 system this ability to move all the way between the two extremes is fully supported. To highlight these two extreme cases we constructed a side-by-side video showing one person using only speech and one person only gesture. Each builds essentially the same structure - and in roughly comparable time. Figure 13 is a screen shot from this video.

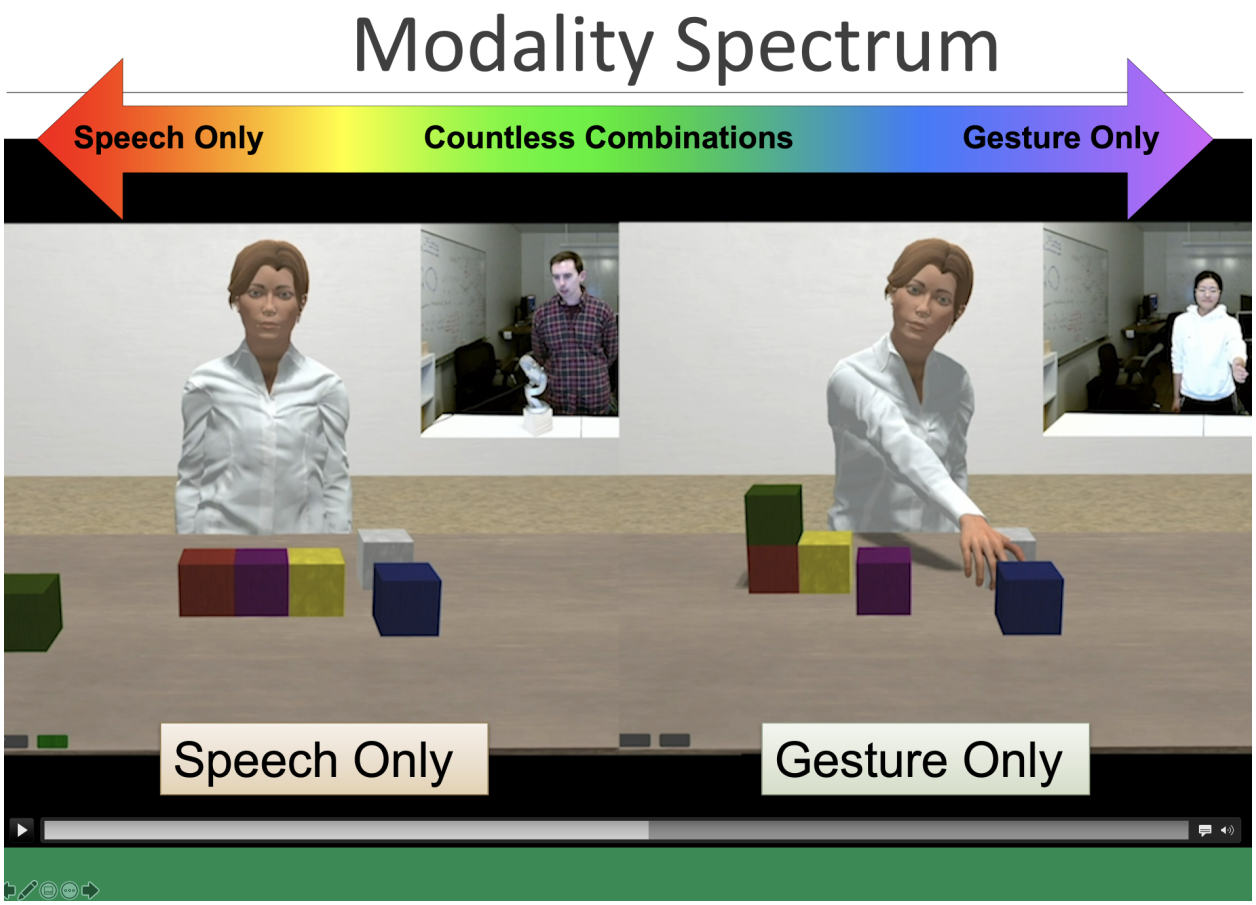


Figure 13: Snapshot from the Modality Spectrum video

Here also is a link to this video.

[Modality Spectrum Video]

What can be observed in this video is a side-by-side comparison of two different people working with the same Diana system building a staircase. The person on the left is using speech alone, the person on the right is using gestures alone. Needless to say the team is proud of the fact that the system is able to behave at both of these two extremes. However, what is actually even more important is that in order to arrive at these two extreme capabilities, it was necessary to build a system with tremendous flexibility relative to how speech and gesture are integrated. To put this another way, there exists a combinatorial explosion of different multimodal styles of interaction with the Dyana system in which individual users may choose to use more or less of one modality relevant to another to match their own preference.

One last complete video highlighting interaction with the Diana 2.0 system is included here. This video runs just over one minute, and still is able to highlight a wide variety of mixed modality interaction. It also illustrates that Diana 2.0 is designed to anticipate a user's wishes and avoid excessive back-and-forth turn taking behaviors.

Diana's World video showing Diana 2.0

Generating and Responding to Facial Expressions

Our team came into the CWC program with 15 years of experience working with human face recognition and that included familiarity with techniques for recognizing different facial expressions. Generally speaking, research along these lines falls under the broad heading of affective computing where the goal is systems that can interpret and respond to different human affects, e.g. happy, sad, etc.

For the first several years of our project, this connection to affective computing was left unexplored while we justifiably pushed hard on the more immediately compelling issues of integrating speech and gesture into what grew into the Diana 2.0 system. However, starting in 2018 a series of experiments were conducted looking at facial expressions as manifest in the signaler and builder data from the eggnog data set. While informative at one level, it's worth noting that even in this early effort, if you can clear that the software we were using to interpret facial expression and weaknesses and just as important, it wasn't terribly obvious that facial expression was playing anything like a principal role in the nonverbal communication between signaler and builder. More about this work can be found in (Wang et al., 2021).

After looking at the eggnog data, our attention shifted to implementing fully by directional affective computing in the Diana 2.0 system. To be precise, this meant building a variant of the Diana system that used off the shelf facial expression

recognition software to constantly interpret the expression of the user. It also included significant modifications to the avatar so that the avatar could generate appropriate expressions in response to the task context and the user.

The term “appropriate“ here is really the crux of the matter. A tremendous amount of research has been done on expression recognition, and a lot of research has been done showing how an avatar’s affect might improve her users impression or satisfaction working with an avatar. What’s rare in the literature, and in truth we are not aware of any prior example, is the empirical study of bidirectional affect when situated in a concrete problem solving environment. Again, to be precise, the goal of a user working with Diana is to build stuff, and it is clear first and foremost a user's attention is focused on the goals of building a structure. It is in this context that we set out to measure if, and if so, how much, support for bidirectional affect might improve a user’s experience.

To tackle this question, we specifically had to implement bidirectional affect and come up with strategies for how an avatar might best respond to a user’s affect. Specifically, we did the following. First, we integrated the Activa (get version details) expression recognition system into the dialer 2.0 architecture. This involved generating a real time face tracking capability tied into the red green blue channel coming off of the Kinect. The face was then fed to the Activa system which would generate a stream of expression labels. The expressions Diana was capable of recognizing were Neutral, Joy and Anger. In addition she interpreted pointing gestures and several forms of negative acknowledgement such as “Nevermind” when considering how to appropriately respond to a user.

The second step was to dive deeply into the Unity software that controls the Diana avatar. It was necessary for us to build from scratch the extra controls that could be used to make Diana smile, or look thoughtful, etc. Specifically, the expressions that Diana was able to exhibit were Joy, Sympathy, Confusion, Concentration and Neutral. It should be noted that Concentration is very important when working to solve problems with a person despite the fact that it is almost never listed among the common affective states.

The third step was actually considering what might be the best way for an agent such as Diana to respond to the affect of a user. This led us to consider three options. The first option is essentially a baseline and it’s the system prior to the inclusion of affect. In other words, Diana‘s facial expression is neutral and remains forever unchanged (Emotionles). Option two is very easy to implement, and it’s simply called mimicry. Diana essentially sets out to mirror the user; if the user smiles, Diana smiles, if the user is unhappy, Diana looks unhappy, etc. It’s not hard to see where the mimicry option

may fall into trouble. If he is or is getting angry, reflecting that anger back at the user is almost certainly not a good idea.

Therefore, the third option is responsive, by which we mean Diana follows a set of rules built to better connect the user with Diana. For example, if the user is smiling, having Diana smile back at them is a great idea. However, if the user looks angry, then Diana adopts a deeply concerned facial expression. For the record, the deeply concerned expression is not generally found described in the effective computing literature, it is really an outgrowth of recognizing that expressions need to match up with human expectations relative to two way communication and a collective problem-solving task.

To measure user response to the three different bidirectional modes of affective computing, we developed an experimental protocol and ran 21 subjects through our experiments. Each subject interacted with the three different Diana variance. Throughout the experiment, subjects were asked to build a block structure working with Diana. After each trial working with Diana the subjects were asked to fill out a brief survey with seven questions calibrated to place responses on a standard likely scale. The subjects were not told anything in advance about the differences in the three versions of the system they were using, in the order the systems were presented to users was randomized.

Figure 14 below shows a summary of the results. There is a lot in this graphic, let's begin the explanation by explaining that along the left-hand side are the actual questions to which the users are responding. In the middle with a color-coded bar chart the proportional responses are shown in such a way that the neutral (Sierra response) six point forms a vertical line. This mode of displaying makes it easy to scan down the right and left to see if there is any difference in proportion of negative versus positive feedback from the user. On the right hand side the three conditions, mimicry, emotionless (neutral) and responsive (demo) are shown in legends in order to make quick scans of the figure easier.

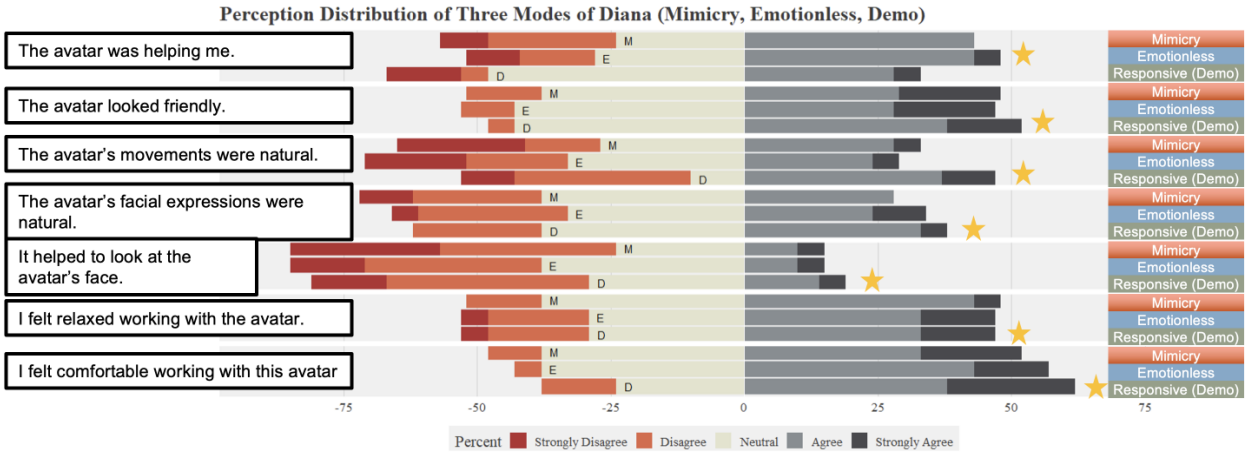


Figure 14: Summary of findings for our bi-direcgtional study of affect.

In terms of building confidence that bi-directional affect with a responsive design is valuable, the strongest statement we can make is that for six of the seven questions the overall most positive feedback is associated with the responsive version of Diana. This is a very encouraging first result. However, it must immediately also be noted that the differences are not very large. It is for this reason that our choice of words has been somewhat careful. We think it's important to continue to work on improving bi-directional affective computing, but we also suspect that affect will never become the dominant factor contributing to overall user satisfaction. Put simply, a naïve user is working with the system in order to accomplish a task, and they are most likely most aware of how well the agent is or is not understanding and responding to their intentions.

2021 Naive User Study

The evaluation of the Diana 2.0 system mirrored that of the Diana 1.0 system with one main addition. At the beginning of the study, participants were instructed to read out loud a paragraph containing the instructions for the study. The recording of the reading was used offline to calculate the word error rate of the speech to text module. We hypothesized that a higher word error rate caused by speech to text would result in lower SUS scores. Thirty participants (15 male, 15 female) aged 18-57 (mean=27, SD=11.8) participated in the study.

Similar to our previous study, we observed a high completion rate with only 24 of 240 tasks not completed (10%). Of these 24 incomplete tasks, thirteen resulted from Diana getting stuck in a state that required the task to be restarted. We noticed that there were situations where Diana was attempting to perform one-shot learning (i.e., learn a new gesture she observed) and was not provided the feedback from the participant she

expected and would result in Diana being stuck in an unstable state. In these cases, Diana would be restarted and often resulted in the participant running out of time for that particular layout.

Evaluation of System Usability Scores showed a significant improvement over the Diana 1.0 system. SUS scores ranged between 67.5-90, with a mean of 74.3 (SD 8.2). Recall that a score of 68 is considered above average and anything above 80 as excellent. We observed only four scores below 68, all of which barely missed the bar with a score of 67.5. We saw eight SUS scores (27%) above 80. From these scores, we can conclude that the Diana 2.0 system made significant improvements in the system's overall usability. Feedback from participants was also positive and highlighted that participants appreciated the multimodal aspect of the interaction. For example, Participant 5546 stated, "the combination of speech and gesture at the same time as useful and unique".

Participants also had some suggestions for future revisions of the system. Multiple participants commented on the commands left, right, in front, and behind. The current system uses itself as a reference when mapping directional commands, for example, "In front of the blue block." This design decision was made based on our initial elicitation studies observing human-human collaboration. However, participants in our study expected the system to use themselves as the reference. This resulted in participants having to redo actions until they were comfortable with the system. There is also feedback that this mirroring may have added to participants' cognitive load. For example, Participant 5919 stated, "[I disliked] having to think of the opposite movement I would [need to] make since the system's perspective was flipped from mine. That took the most getting used to."

Another common complaint from participants was the recognition of speech commands, in that they felt that at times Diana did not correctly interpret their commands. This was a concern for us as we developed Diana using the Google Speech-to-Text API as we noticed that it sometimes had difficulty with non-Western accents. Therefore, we calculated the word error rate (WER) using the paragraph read by each participant at the beginning of the study as an indirect measure of the Speech-to-Text API during the study. When correlating these values with participants' SUS scores, we found no correlation between the two. Thus, we have no evidence to support our hypothesis that errors in speech-to-text negatively affected system usability scores.

Overall, results from our user study show that Diana 2.0 can be considered a success. Participants continued to be able to successfully collaborate with Diana to create block layouts at a high rate. More importantly, SUS scores dramatically improved over the first version of the system and are now average in the usable range. Future revisions of the

Diana system will examine the mapping of directional commands and continue to the recognition of speech input.

Faelyn Fox

In order to advance one branch of our work supported by DARPA's Communicating with Computers (CwC) Program to better attract external funding (e.g., NIH, NSF) from agencies as CWC was closing, DARPA provided bridget funding to develop and research a transition project, called Faelyn Fox. The team included Dr. Ross Beveridge and Dr. Francisco R. Ortega while adding two new team members, Dr. Lisa Daunhauer and Dr. Anita Bundy. They are experts in developmental assessment and joined our team to help us collectively identify key opportunities where an embodied agent able to see and interact with a person, in this case a child, might provide more reliable developmental assessment data. Faelyn Fox directly builds upon the avatar embodiment, simulation, and visual recognition components of what we and Brandeis University collectively describe as the Diana System (McNeely-White et al., 2019).

Faelyn Fox system was developed to run over a browser using similar yet simpler architecture from Diana. For example, Faelyn fox does not have gesture recognition built-in or any language recognition as it is with Diana. The purpose was to be able to learn from Faylyn Fox as we learned in the early days with Diana, by using a participatory design (i.e., a Wizard-of-Oz). This allows the Wizard to create magic for the children accepting any type of interaction.

Faylyn fox development followed recommendations learned during Diana's project and ones offered by Dr. Daunhauer and Dr. Ortega. Faelyn Fox has an experiment module that allows to run mini-trials of different tasks seeking either developmental and executive functions data collection or gesture elicitation studies for children. The recommendations of Dr. Daunhauer were critical for the right language to use for Faelyn when speaking to children, including requests and feedback. The population we targeted were children between the age of 3-7 years old.

We conducted the research using Azure Kinect to make a videorecording of the participants as they performed their gestures, as this records skeleton data, depth and optical data. In addition, the experiment used a microphone to align the recording of the Kinect without screen recording using OBS software for additional observational analysis. The experiment had the following tasks:

Sort Toys Task: The objective is to examine an element of executive function known as working memory which is the ability to hold limited amounts of information for use in the moment (e.g., doing mental math). In this scenario data collected utilized in this paper asked the participant to put different toys in two different boxes, as shown in Figure 14. There was one small box and one large box. The animals were large or small. At first the children had the task of putting the large animals in the large box and the small animals in the small box. This task was then made more complex after 6 trials with a rule change, when it asked the participants to put the large animals in the small box and the small animals in the large box. The rule change requires the participant to use working memory to hold the rule and use it when applicable during the trials.

Figure 14: Sorting toy by size.

Grooving Tasks: The objective of this task is to examine inhibitory control which is the ability to control one's body or actions to inhibit an automatic, or prepotent response. Faelyn's full avatar is displayed with a set of music speakers to the side. The participants are asked to dance to music anyway they want. After a set period of time the music stopped. The participants were tasked to stop and freeze when the music stopped. The duration of music was 7, 10 and 12 seconds presented randomly. A screenshot of this task is shown in Figure 15.

Figure 15: Grooving Task

Go-No Go Tasks: Two types of boxes are displayed, one with a circle on it one with a star on it. The 2 types of boxes appear one at a time. The participants were asked to point/gesture to the box with a circle on it and not point to ones with a star. The objective of this task is to examine both working memory and inhibitory control. Additionally, the response time for this task is a classic measure of executive function. A screenshot from the Go-No Go task is shown in Figure 16.

Figure 16: Go-No Go Task

Gesture Elicitation Tasks: One of the critical components for Faelyn Fox to be successful is to understand how children perform gestures. In particular, how different ages affect those gestures. A gesture elicitation task, which is a form of participatory design study, was developed to see what were the most common gestures.

This scenario consisted of 6 different tasks. The participants were asked to produce gestures for referents provided verbally by Faelyn. This being a WoZ study, any and all gestures were accepted. The 6 tasks were:

Rotate: (Figure 17a) Three toys in a line were shown to the participants. The toys on either end were facing the center toy. Participants were asked to make a gesture to rotate the toy in the direction of either of the 2 toys, floor, or ceiling.

Scale: (Figure 17b) Single toy is placed in the center of the table. The participants were asked to make gestures to shrink or enlarge the toy to fit into a rendered box.

Translate: (FIgure 17c) A single toy is placed in the center of the table. A rendered sphere, 1 step distance in the front, back, left, right, top, or bottom of the toy appears randomly. The participants were asked to make gestures to move the toy in the sphere.

Create: (Figure 17d) The participants were asked by Faelyn to make gestures to create a toy in a rendered sphere.

Select: (Figure 17e) The participants were asked by Faelyn to make gestures to select a toy among the toys that were created previously.

Remove (Destroy): (Figure 17f) The participants were asked by Faelyn to make gestures to remove a toy among the toys that were created previously on it. When a participant points to a correct box an animal toy pop out.

Figure 17a: Screenshot of the rotate an object elicitation task

Figure 17b: Screenshot of the scale and object elicitation task

Figure 17c: Screenshot of the translation an object elicitation task. .

Figure 17d: Screenshot of the create a new object elicitation task.

Figure 17e: Screenshot of the select an object elicitation task. .

Figure 17f: Screenshot of the Remove (i.e. destroy or delete) an object task.

Experimental Procedure

These experiments were done one-on-one with a single participant and the researcher. The researcher would start all recording devices then begin the FoxWorld application controlling the application behind the participant, without the participant's knowledge. The participant completed multiple tasks listed above at random to avoid ordering effects and user fatigue effects. At the end they were asked their overall feedback and if they had any questions. Due to covid, a small set of participants were able to be recruited. With vaccines recently released for children between 6-11 and with the hope that vaccines for kids between 3-5 become available soon, we expect to be able to recruit more participants. Therefore, these results are only preliminary.

Preliminary Results

Sort Toys Task: Majority of the participants pointed with their index finger of dominant hand. One participant made a swiping gesture. One participant tried different gestures as trials went on. Most of the participants gestured to correct boxes, only one participant had to be reminded of the rule switch after 13 trials. Most participants showed medium to high motor skill when gesturing.

Grooving Task: Most of the participants followed and froze at the correct time. Only one participant did not freeze for one trial. They corrected themselves in the next trial. The response time to freeze ranged from 0.5 seconds to 2 seconds.

Go-No Go Task: Majority of participants made pointing gestures to select the circle box and none of them selected the star box (which was the incorrect response). There were 4 instances of self-correction where they almost pointed at thestar box but retracted their hands. The response time for selecting the correct box ranged from 0.5 seconds to 1.3 seconds.

Gesture Elicitation Task: For rotation, majority of the participants swiped in the direction of the rotation, some just pointed where they wanted the toy to look. In case of scaling, there were 3 different types of gestures presented. Swipe, point, and whole hand pinch and spread for shrink and enlarge. For the translation task, most of the participants either pointed to the location or swiped with the index finger of the dominant

hand in the direction of the destination location with one participant using their whole palm to swipe. Creation was the task where the participants hesitated a bit in the beginning trying to think of a way to make a gesture. Majority just pointed to the location. One participant spread the finger proportional to the size of the toys differentiating smaller and bigger toys. As expected, selecting an object produced the most consistent gesture of pointing to the object. In case of removing toys, the participants made swiping gestures with index finger (diagonal or horizontal) or pinched fingers (away from table). This being the last task in elicitation, some of the participants realized that no matter what gesture they do, it will be recognized so 2 participants kept changing the gestures they performed.

Faelyn Fox Project Next Steps

Faelyn Fox data collection will be completed as soon as more children will be available. It is expected that at least 2-3 publications come out of the studies. The next step will be to go to funding agencies where Dr. Daunhauer will lead for NIH and Dr. Ortega for NSF. In addition, with the results at hand, a more automated system will be able to be developed.

Project Publications

The following publications are direct products of this project.

Nikhil Krishnaswamy, Ross Beveridge, James Pustejovsky, Dhruva Patil, David G. McNeely-White, Heting Wang and Francisco R. Ortega, “Situational Awareness in Human Computer Interaction: Diana's World”. (2020). In *International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments*, December 4 2020. (Short paper and live demonstration). (Best Demo Award)

Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge and James Pustejovsky, “Diana’s World: A Situated Multimodal Interactive Agent”. (2020). Proceedings of the AAAI Conference on Artificial Intelligence , April 3, 2020, Vol 34, Issue 9.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Bruce Draper and James Pustejovsky. “Communicating and Acting: Understanding Gesture in Simulation Semantics”. (2017). In *12th International Conference on Computational Semantics*, Montpellier, France, Sept. 19 – 22, 2017.

David G. McNeely-White, Francisco R. Ortega, J. Ross Beveridge, Bruce A. Draper, Rahul Bangar, Dhruva Patil, James Pustejovsky, Nikhil Krishnaswamy, Kyeongmin Rim, Jaime Ruiz, Isaac Wang, “User-Aware Shared Perception for Embodied Agents”. (2019). In First IEEE International conference on Humanized Computing and Communication (HCC 2019), September 25 - 27, 2019, Laguna Hills, CA.

Pradyumna Narayana, Ross Beveridge and Bruce Draper, “Gesture Recognition: Focus on the Hands”. (2018). In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Salt Lake City, June 19-21, 2018.

Pradyumna Narayana, Nikhil Krishnaswamy, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Kyeongmin Rim, Ross Beveridge, Jaime Ruiz, James Pustejovsky, and Bruce Draper, “Cooperating with Avatars Through Gesture, Language and Action” (2018). In IEEE Intelligent Systems Conference (IntelliSys), London, September 6-7, 2018, pp. 156-165.

Heting Wang, Vidya Gaddy, James Ross Beveridge and Francisco R. Ortega, “Building an Emotionally Responsive Avatar with Dynamic Facial Expressions.” (2021). In *HumanComputer Interactions*”. In Multimodal Technologies and Interaction, 5:3

Isaac. Wang, Mohtadi Ben-Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge and Jaime Ruiz. “EGGNOG: A continuous multi-modal data set of naturally occurring gestures with ground truth data”. (2017b). In *IEEE Conference on Automatic Face and Gesture Recognition*, Washington DC, May 31 – June 2, 2017

Isaac Wang, Pradyumna Narayana, Dhruva Patil, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. (2021). It’s a Joint Effort: Understanding Speech and Gesture in Collaborative Tasks. In *Human-Computer Interaction. Interaction Techniques and Novel Applications (Lecture Notes in Computer Science)*, 159–178. https://doi.org/10.1007/978-3-030-78465-2_13

Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. (2017a). 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)*. ACM, New York, NY, USA, 2990-2997. DOI: https://doi.org/10.1145/3027063.3053239

Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge and Jaime Ruiz. “Exploring the Use of Gesture in Collaborative Tasks”. (2017c). In *ACM CHI Extended Abstracts*, Denver, CO, May 6 – 100, 2017

Isaac Wang, Pradyumna Narayana, Jesse Smith, Bruce Draper, Ross Beveridge and Jaime Ruiz, “EASEL: Easy Automatic Segmentation Event Labeler”. (2018). In *ACM Conference on Intelligent User Interfaces*, Tokyo, March 7-12, 2018.

Other References

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, *12*(2), 213-253.

Krishnaswamy, N., & Pustejovsky, J. (2016, December). VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 54-58).

Strout, J. J. (2020). *Multimodal Agents for Cooperative Interaction* (Master's thesis, Colorado State University).