Sharing A Single Expert Among Multiple Partners

Jeffrey Wong¹, Lui Min Oh², Jiazhi Ou¹, Carolyn P. Rosé¹, Jie Yang¹, and Susan R. Fussell¹

¹Carnegie Mellon University

5000 Forbes Avenue

Pittsburgh, PA 15213

[jeffwong, jiazhiou, jieyang, sfussell]@cmu.edu

cprose@cs.cmu.edu

²DSO National Laboratories

20 Science Park Drive

Singapore 118230

keithoh@dso.org.sg

ABSTRACT

Expertise to assist people on complex tasks is often in short supply. One solution to this problem is to design systems that allow remote experts to help multiple people in simultaneously. As a first step towards building such a system, we studied experts' attention and communication as they assisted two novices at the same time in a co-located setting. We compared simultaneous instruction when the novices are being instructed to do the same task or different tasks. Using machine learning, we attempted to identify speech markers of upcoming attention shifts that could serve as input to a remote assistance system.

Categories and Subject Descriptors

H5.3. [Information interfaces and presentation]: Group and Organizational Interfaces – collaborative computing, computer-supported cooperative work.

General Terms

Experimentation, Human Factors.

Keywords

Remote collaborative physical tasks, video-mediated communication, remote expertise, multi-party conversation, focus-of-attention, multi-tasking, grounding

INTRODUCTION

As technology becomes more complex, the expertise needed for its operation and repair becomes both more essential to the daily operation of many organizations and more limited in supply. Work requiring specialized knowledge may be completed more efficiently and less expensively by novice workers who are assisted by a remote expert via information technologies. For example, it can be faster and cheaper to provide technical support over the phone than to travel to a customer's site to fix a problem directly. Expertise thus becomes a limited resource that is allocated across novices in diverse locations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

Copyright 2007 ACM 978-1-59593-593-9/07/0004...\$5.00.

In this paper, we focus on how experts allocate their time and attention across multiple novices who are engaged in *physical tasks*, such as aircraft repair or surgery, in which individuals manipulate 3D objects in the real world. Previous studies ([9][15][16]) examining one-on-one assistance in physical tasks have identified communication features that are essential for remote instruction. For example, instructional dialogues are more efficient when the expert can see the novice's workspace via video conferencing technologies [15] and when the expert can use a pointing tool to indicate objects and locations in the workspace [10].

Figure 1. A single remote expert providing guidance across multiple worksites

While the systems described above are effective for one-on-one remote assistance, they do not support situations where a single expert helps multiple novices in different locations. Consider the scenario illustrated in Figure 1 where a remote expert is provide expertise for aircraft maintenance at multiple sites. Unlike in the one-on-one case, the personnel at each site cannot assume that they have the expert's undivided attention. If the audio channel is a party-line, on-site personnel may be uncertain about which of the expert's instructions is intended for them or others. Although remote personnel could be helped in sequence, there may be periods when the expert is waiting for the remote personnel to perform certain

instructions. This time could be better used to assist other personnel at other sites.

With modern information technology, an expert should be able to assist multiple workers in parallel. An expert could monitor video feeds from different novices' workspaces and direct his/her attention to a novice on an as-needed basis. A system of this type could be of substantial value, especially in time-critical settings such as the disaster relief scenario illustrated in Figure 1.

The features required by a system to support parallel remote multi-party instruction are not easily determined from research on one-on-one assistance. New issues of situational awareness [8] and conversational grounding [6] arise when an expert must keep track of multiple workspaces, only one of which is the current focus of attention. For example, the expert may miss critical events occurring at non-focal sites. Novices may be unsure of whether their actions are being observed; if they don't hear corrective feedback from the expert and they believe they are being watched, they may assume that their actions are correct [11]. Because simultaneous instructive collaboration over multiple video channels is not something that has been widely investigated, we wanted to conduct a face-to-face study of simultaneous instruction examine these issues before designing an electronic prototype.

Another issue is how the system supports the focus of the expert. One possible system design might assume that an expert would have ample screen real estate to be able to see all video feeds at the same time. However, there are good reasons for a system to support focusing on a particular feed and switching focus among novices. Network bandwidth can be limited where the experts, especially those who telecommute from a residence. With their GAZE-2 system, Vertegaal *et al.* [22] observed that bandwidth usage can be reduced if a system reduces the video quality of unattended video feeds.

A notion of focus allows the expert's speech to be directed towards one novice and not others. This could help novices know when they are being addressed. Audio attention also works in the reverse direction. GAZE-2 supported an "artificial cocktail party effect" where the audio from unattended locations is lowered so the video feed in focus is loudest. In a remote repair context, ambient noise from environment of unattended feeds could be distracting.

Manually operating camera controls can, however, reduce task performance. Ranjan, Birnholtz, and Balakrishnan [19] created a Wizard-of-Oz experiment comparing manual pan-and-zoom against a simulated, automatic control system. Using features of the task and the helper/worker dialog, they showed that subjects performed better with automatic pan-and-zoom than when they controlled pan-and-zoom manually. They posited that the extra effort of using the camera controls may have discouraged users from using them at crucial points in the tasks. Our prior work indicated

that the desired focus-of-attention in a remote expert can be predicted by using properties of the task, the novice's actions, and the spoken dialog.

Although switching between views of novices is different, automated switching could be useful in our remote repair scenario. Deciding whom to switch to can be a burden for the expert. The expert would first have to determine who is idle. If fairness is important, the expert needs to remember who hasn't helped recently. Idleness and fairness are variable of the remote work sessions that a system could track while the expert's focus is on individual novices. Also, if a novice is confused or has an emergency, the expert's attention can be diverted to that novice without the having expert search the other video feeds for source of the request. Therefore, a secondary goal of this study was to identify speech and visual cues that could be used as the basis for intelligent, automatic switching of attention.

In the rest of this paper, we describe the theoretical motivation for our work, identify the research issues in our experiment, and present our qualitative and quantitative analyses of experts' multi-tasking behavior. We conclude with a discussion of the implications of our findings.

THEORETICAL BACKGROUND

The foundation for our work is based on theories of conversational grounding [4][5] and joint projects [4] developed by Clark and his colleagues. In the remainder of this section, we briefly introduce the concepts of grounding and joint projects and describe how they apply to multiparty assistance dialogues.

Grounding

Conversational grounding refers to the interactive means by which participants ensure that messages have been understood as intended [5]. When people speak, their utterances are based on what they believe their listeners know. After speaking, they look for evidence of the listener's understanding. If the utterance appears to have been understood, the information enters the common ground between the speaker and the listener. If the listener displays evidence of misunderstanding (e.g. a perplexed look), the speaker will provide clarification.

Conversational grounding in dialogues in which an expert guides a novice on physical tasks tends to follow a predictable pattern [15]. First, collaborators come to mutual agreement upon, the objects to be manipulated using referential expressions. These objects are "grounded." Next, they provide instructions to be performed on the objects. Finally, they check status of the task to ensure that the actions have been understood and carried out correctly. They use two important resources in this grounding process: the discourse history and the physical space shared by communicators [4]. When collaborators share a physical setting, an expert can refer to objects by pointing and a novice can evidence his or her understanding by picking up the correct object [11]. Without visual information,

collaborators must use lengthier verbal descriptions of objects and acknowledgements of understanding.

Grounding in remote multi-party conversations may differ from grounding in two-party dialogues in several ways. Novices cannot be sure that the expert is watching their actions and that those actions have entered the common ground. When a novice is the only one being helped, he or she can interpret lack of corrective intervention by the expert as acknowledgement that his or her actions are correct. With more than one novice, a lack of intervention is ambiguous—either the expert is watching and the behavior is correct, or the expert is not watching. Also, if an expert issues an instruction to a novice but does not wait for evidence of grounding before switching to another novice, he or she may need a update of the current task state when he or she returns to assist that novice.

Joint Projects

Clark [4] has argued that conversation can be conceived as a “joint project” comprised of three stages: an *entry*, a *body*, and an *exit*. The concept of joint projects applies to conversation at several levels of granularity. At the broadest level, an entire conversation can be thought of as a joint project, from the initial greeting (e.g., “Hi, how are you?”) through the body of the discussion, to an exit (e.g., “Well, I have to get going”). Smaller units of conversation can also be considered joint projects, embedded in a larger joint project. For example, the exit phase of a conversation may itself consist of an entry (hinting at goodbye), a body (an exchange of goodbyes), and an exit (walking away).

Bangerter *et al.* [2] found that participants in certain task-oriented conversations used different speech markers to indicate transitions between joint projects and continuations within them. Phrases such as *uh-huh*, *yeah*, and *right* indicate the intent to continue with the current project while phrases such as *okay* and *all right* indicate shifts to a new joint project. Conversational partners use these markers to keep track of where they are within the dialogue structure.

The concept of joint projects is a useful way to understand the brief conversations that an expert has with each novice in multi-party help dialogues. In these conversations, the expert instructs one novice for a period of time and then switches to another novice. Continuous dialogue with one novice can be seen as one joint project (which may itself include more than one subproject) while a switch in attention is a transition to another joint project. Drawing inspiration from Bangerter *et al.*, we hypothesized that experts might naturally use speech markers to signal to novices that a joint project is continuing, that a joint project is ending, or that a new joint project is about to be initiated.

Goals of the Current Study

We conducted a study in which a single expert (*Helper*) assists two novices (*Workers*) as they individually construct sections of a large toy robot. We consider two scenarios, one in which all novices are performing the *same* task

(similar to students being guided by a lab instructor) and the other in which all novices are performing *different* tasks (similar to the typical remote assistance scenario).

The primary goal of the current study was to understand how experts allocate their visual attention and instruction-giving over multiple novices in different locations. Although prior work has examined the dynamics of multi-party conversations (e.g., [1][13][21]), to our knowledge, our work represents the first attempt to understand a conversational scenario in which multiple conversations about 3D objects occur, but there is overlap between participants across conversations (but see [1] for a study of parallel conversations in an audio space and [12] for a discussion of parallel conversations in online chat.)

Our more specific aims were threefold: First, we wanted to analyze qualitatively the ways that experts allocated attention across multiple novices, with a specific focus on the management of joint projects.

Second, we wanted to determine whether we could identify markers delimiting the boundaries of joint projects as Bangerter *et al.* did for telephone conversation. If markers of the beginnings and endings of joint projects with novices can be identified, video systems can use these markers to switch experts’ attention automatically as a function of task progress. To this end, we used machine-learning algorithms to analyze the relationship between experts’ speech and their shifts in visual attention.

Third, we wanted to explore how experts allocated attention across multiple novices when these novices were performing different tasks and when they were performing the same task (individually). In the *same task* case, a single expert instructs two or more novices on the same task at the same time. This situation is akin to certain types of classroom instruction, such as laboratory courses in science. In the *different task* case, the expert instructs two or more novices on different tasks. This situation is akin to tech support hotlines. We analyzed how much time experts spent helping a novice before switching to another novice, what events prompted attention switches, and whether speech markers could reliably predict an upcoming switch.

METHOD

Design

Seventeen groups of three participants (one Helper and two Workers) completed robot assembly tasks. Each group completed one task in the *same task* condition, in which both Workers assembled the same part of the robot, and one task in the *different task* condition, in which each Worker completed different tasks. Helpers wore eye-tracking equipment so that we could determine which of the two workspaces they were looking at as they spoke. The order of conditions was counterbalanced across groups. Sessions were audio/video recorded and transcribed.

Apparatus

Workers sat at tables on opposite sides of a tall barrier that divided them. The Helper sat at the side of the junction of the two tables with the barrier dividing his or her field of vision. Each workspace was covered with a different colored paper (red or green) in order to allow us to code the Helper's gaze by applying machine-learning techniques to the head camera video. Figure 2 shows the table from the Helper's point of view.

Helpers wore an ISCAN head-mounted eye-tracking system that recorded their gaze overlaid onto a video recording of the field of view. An additional camera was placed to the left of the Helper to record the Helper's gestures and body movements. Helpers and Workers wore wireless microphones whose output was mixed into a DV-VCR as the audio track for the head camera.

Figure 2. The view of the Workers's work spaces from the Helper's perspective

Tasks and Materials

Tasks

The Robotix Vox Centurion robot kit (Figure 3) was used as the basis for three tasks of similar difficulty, each approximately 15 steps and 10 minutes long. Each task resulted in the completion of one part of the robot: the right ankle, left foot, and left arm. An instruction manual outlined the steps to be completed.

Surveys

Participants completed four sets of surveys. A brief *pre-test survey* collected basic demographic information. *Post-task surveys* administered after each task asked about how well the group worked together, confidence in completing the task correctly, fairness in allocation of Helper attention to each worker, and assessments of the Helper's ability to multi-task. A *final survey*, completed at the end of the experiment,

Figure 3. The robot used in the experiment.

contained general questions about participants' experiences.

Participants and Procedure

Participants consisted of 51 undergraduate students who were screened for colorblindness. They were paid \$15 for their efforts. Participants were run in groups of three. One group member was randomly assigned to the Helper role, while the others were Workers. The worker on the Helper's left was the "green" worker and the other was the "red" worker. Groups were told there was a \$25 at the end of the semester for the fastest and most accurate group.

Participants first completed the pre-test demographic questionnaire. The Helper was then fitted with the eye-tracking headgear and the equipment was calibrated. Next, Workers were seated at their respective tables with unassembled robot pieces in front of them. Helpers were instructed to assist both Workers at the same time. Helpers were told not to touch the parts in the workspaces.

In this study, we used trained novices to be Helpers, based on prior research that found no difference in the amount or content of assistance provided by experts vs. trained novices for certain tasks [16]. Before instructing Workers, Helpers were moved out of Workers' view and assembled the robot parts according to instruction sheets for the tasks. These instruction sheets were also available to Helpers as they guided the Workers.

The group then performed their assigned tasks in one condition (either the *same task* or *different task* condition) and then remaining tasks in the other condition.

Transcribing Conventions

Conversations were transcribed in detail, including pauses and other disfluencies. Using the videos, we then selected a subset of exchanges for more detailed transcription of body position, eye gaze, and speech overlap. For the purposes of readability, we present transcript excerpts in standardized English rather than phonetically. In the excerpts, the Helper is denoted by *H*, the worker at the red table is denoted by *W1*, and the worker at the green table is denoted by *W2*. Actions and nonverbal behaviors are indicated by italicized text within triangular brackets. Dashed horizontal lines are used to indicate beginnings and endings of joint projects between the Helper and a single Worker.

Data Collection and Analysis

We used machine-learning techniques to analyze the video from the Helper's head-mounted camera to automatically determine who the Helper was looking at for each frame of video. As shown in Figure 2, each Worker sat at a different colored table. The parts of the tables that were closest to the Helper were covered in blue, to classify the gaze when the Helper was looking the manual placed on his or her lap. We then used the proportion of red, green, and blue in each frame as features. We built a decision tree classifier that takes these features as inputs and determines, for each frame, whether the Helper was looking at the Worker at the

red table, the Worker at the green table, the instruction manual, or none of these. The accuracy was 92.7% on a test set of 400 examples. We then ran the classifier on all frames and outputted the gaze coding at each time point. The classifier and human coding of the frames exhibited strong agreement ($\kappa = 0.89$).

We used manually transcribed transcripts to identify the words in the dialog. To find the start and end times of each word, we fed the transcripts into a speech recognizer that found the location of each word in the audio tracks. The times were precise to 10 milliseconds.

RESULTS

We discuss the findings in three parts. First, we discuss joint projects and how they differed in the same task and different task conditions. Then, we show how we attempted to use machine learning to determine whether a set of markers can be used to reliably identify the beginning and end of joint projects. Third, we discuss some of the Workers' behaviors that influenced the Helpers' multi-tasking strategies.

H:	And now take a um the three prong black piece.
H:	<Looks at W2>
H:	You were just holding it right there.
W2:	Ok.
W2:	This one?
	<Looks at W1>
H:	Ok
H:	and have it so it's facing upward- upward.
H:	<Looks at W2>
H:	Like that.
H:	<Looks at W1>
H:	You both got it.

Figure 4. Fine-grained interaction in the same task condition.

Joint Project Structures

As discussed earlier, instructional dialogues for collaborative physical tasks have a regular pattern consisting of establishing reference to task objects, providing instructions for those objects, and ensuring that the instructions have been understood correctly [10]. Within the theoretical framework of joint projects, each phase of the dialogue, as well as sub-sequences within each phase, can be considered a joint project. In a general sense, our multi-party instructional dialogues fit this pattern of embedded joint projects. However, some transitions between joint projects occur *within* speech to one Worker whereas others occur in transitions *between* workers. The strategies used to initiate and continue projects differed in the *same task* and *different task* conditions.

Same Task Condition

The structure of joint projects in the *same task* condition varied between Helpers. Some Helpers were careful to start

by asking their workers to orient their work pieces in the same direction. This allowed the Helpers to issue the same instructions to both Workers in parallel, with minimal amounts of Worker-specific grounding. Helpers who didn't use this strategy had to provide specific information (e.g., describing the correct orientation of a part) to each Worker after telling both Workers where to put the part. One Helper worked with the Workers individually and addressed few utterances to both Workers, but this strategy was very rare in the *same task* condition.

Helpers differed in the granularity of the instructions they provided. Some Helpers issued very detailed instructions. In Figure 4, the Helper first tells the Workers to pick up a piece, checks that they have selected the correct pieces, instructs them to rotate it "upward," and verifies that the piece positioning is correct.

When Helpers addressed both parties simultaneously, they engaged in a joint project involving three people. To repair misunderstandings with an individual Worker, Helpers would enter a sub-project with that Worker (a *side-sequence* in Jefferson's terms [14]) and then shift back to group conversation once clarification had been achieved. Helpers sometimes did this by moving their gaze rapidly between Workers or by explicitly addressing both Workers (as in the last utterance of Figure 4). Although instructions were often addressed to all, individualized clarifications after each instruction or set of instructions were common.

Different Task Condition

In the *different task* condition, Helpers assisted one Worker for about 30-60 seconds before switching to the other Worker. They tended to switch from one Worker to the other when the current Worker was occupied with a compound task, was busy undoing a mistake, or had received enough attention.

H:	And you're gonna connect the small black piece, one, two rotating pieces, small black pieces.
W2:	These?
H:	To the third hole towards the bottom on both of those.
W2:	Ok. These two? The ones that rotate round?
H:	Yeah.
H:	<turns towards W1>
H:	And you're gonna take the--
W2:	Wait, hold on a second. I-- Attach them where?
H:	<turns towards W2>
H:	Ok. To the, the third--

Figure 5. Preempted Helper switch in different task condition.

Although Helpers typically gave a Worker an entire step before switching to the other Worker, they often did not wait to establish that an instruction had been carried out correctly. This can be seen by the many cases where a Worker who had received instruction before a switch would ask the Helper (now instructing the other Worker) to return immediately and provide clarification. In Figure 5, the

Helper gives instructions to W2, turns to W1, and begins instructing W1. However, W2 realizes she doesn't understand her instructions so she interrupts the Helper immediately to seek clarification.

Gaze and Joint Projects

Since we were interested in attention allocation, we wanted measure the length of Worker-level joint projects by using gaze as a proxy for joint projects with each Worker in the same and different task conditions. Gazes at the manual, where frames showed mostly blue paper, were classified as assistance to the Worker who was looked at next.

An analysis of the automatically classified head camera video showed that gaze periods in different task condition were significantly longer than those in the same condition ($t(11) = 6.34, p < 0.001$). Figure 6 shows gaze shifts between Workers in the same task and different task condition for one representative Helper.

Figure 6. Gaze shifts between Workers for one representative Helper across a 300-second interval.

We observed that gaze shifts approximated the beginnings and endings of joint projects in the different task condition. There, the Helper gazed at a Worker while engaging in a joint project with him or her. When the Helper changed addressee, his or her gaze shifted to the new Worker.

In the same task condition, instructions were often issued to both Workers and the Helper would gaze back and forth between Workers. Thus, gaze shifts were not reliable indicators of changes in addressee for that condition.

Helpers' Speech Markers

We hypothesized that we could identify speech markers indicating shifts between joint projects as Bangerter et al. [2] did. We were particularly interested in identifying entry and exit markers of joint projects with different addressees, because such markers could serve as input for an automated camera system. Entry markers could be used to determine if a switch *has* happened, and exit markers could determine if a switch *will* happen shortly. Because, as we noted, joint projects in the same task condition often involved both Workers, we focused on identifying speech markers that delimited joint projects in the *different task* condition.

Our initial approach to identifying entry and exit markers involved a corpus analysis of the transcripts collected in the *different task* condition. We first segmented each transcript into joint projects using our automated gaze analysis to

identify places in the discourse where switches in the Helper's gaze were found. We then examined the regions surrounding those boundaries in the transcripts to identify candidate discourse markers that might reliably predict the Helper's change of attention. Specifically, we looked at the distribution of task-independent markers that might be sufficiently general to be useful in an automated system. These include: *ok*, *yes/yeah*, *uh*, *going/gonna*, and *um*. We broke each joint project into deciles based on its length and calculated the distribution of the markers across the deciles (Figure 7).

Figure 7. Distribution of markers across the 10 deciles within joint projects in the different task condition.

As shown in Figure 7, we found that discourse markers often indicated entries to joint projects but not exits from joint projects. For example, *Ok* and *going/gonna* were more frequent in the first decile than the other deciles, but there was no single-word marker that was found to be highly indicative of the exit of a project. *Yes/Yeah* had the highest frequency in the last decile, 15%.

Some Helpers used of markers very consistently while others rarely used them at all. Furthermore, while the “OK” marker was significantly more likely to be the entry of a joint project with a new worker, “OK” was often a marker for sub-projects within a conversation with one Worker.

Because the markers from our initial corpus analysis could not reliably identify shifts in joint projects, we tried a machine learning approach that uses a wider variety of features that are automatically extractable from the transcripts. Distinguishing samples of conversation that occur on a boundary between joint projects from those that are within a joint projects is very difficult; in part because the distribution of instances that are boundaries versus those that are not is so heavily skewed; and in part because, in reality, joint tasks are closed gradually over a span of time that is much broader than what is captured by a single marker. Thus, rather than setting up the problem as a binary classification problem where we distinguish instances

located at the same time as a shift in joint projects from instances located within a joint project, we instead set up the problem as a numeric prediction problem where, at each instance within a joint project, we predict how close we are coming to the completion of the joint project.

We set up our data so that the classifier could be trained to detect progress in the conversation. We included several new features. For example, rather than only including certain markers we identified in our corpus analysis, we added binary term vectors to encode a representation of the entire discourse that had transpired since the last shift, separated into three regions, one for each of the conversational participants. We also included temporal features. Thus, we include three general feature types:

Local features. These are features that directly relate to the current word, including: whether the current word is a speech marker (e.g., *yeah, ok, now, then, and*), the speaker, the amount of silence before and after the word.

Short-term features. These are features extracted from a 10-second window before the end of the current word: number of Helper words, Helper acknowledgements, worker OK's, words from the other worker, words that begin an instruction (e.g., *you'll be, and then, now*), presence of going/gonna or "um", and amount of silence.

Long-term features. These are features extracted from the entire span of time since the last attention shift, including: total elapsed time as well as the accumulated texts from the helper and two workers. Note that these features are really sets of binary features, each representing whether or not an associated term occurred at least once.

We used a machine learning algorithm to predict the distance of each word from a shift in joint projects. We set up the training data by associating each instance with a number between 1 and 15 indicating the distance, in terms of number of instances, to the next instance where an attention shift occurs. We capped the distance measure at 15 since, in practice, the signs of a joint project closing occur relatively close to a shift. In order to emphasize the importance of recognizing the final segment of a joint project span, rather than using this integer value between 0 and 15, the final distance score assigned to each instance was log base 15 of this number, except in the case of a 0 distance, in which case we assigned 0.

We obtained 2259 instances. Using an algorithm for training a support vector regression model [20], we used a technique called 10-fold cross-validation to evaluate the performance of our classifier. In other words, we divided our data into 10 train/test pairs where the training set has 90% of the data, and the test set has 10% of the data, so that we could test each test segment with a model trained on data not included in it. Support Vector Machines (SVMs) have been proven very powerful for numeric applications [3]. Unfortunately, the correlation coefficient between the predicted values and the true values was 0.46 ($p < 0.0001$),

which shows the closeness of a word from the boundary of a joint project is only moderately predictable given our feature set.

Worker Behaviors

In this section, we report some observed Worker behaviors that influenced how Helpers allocated their attention.

Attention Requests

We classified Workers' utterances as attention requests when requests were made when the Helper was looking at the other Worker. Attention requests are questions or signals intended to get the Helper to help that Worker. As with the speech marker analysis, attention switching was so fast in the same task condition that we decided not to attention requesting techniques in that condition.

We observed three main classes of attention requests by Workers: interruptions as the Helper was talking to another Worker, requests made as the Helper was not talking but might have been watching the other Worker, and preempted Helper attention shifts. Preempted Helper shifts are those where the Helper has finished helping the Worker for some period of time and has begun shifting to the other Worker, but the previous Worker preempts this shift by asking the Helper to clarify instructions. We observed that some but not all of these preemptions were actually needed in order for the instructions to be correctly understood.

We identified and classified attention requests by watching the videos and classifying each request according to the above criteria. There were 12 worker requests in our data altogether. Table 1 shows the breakdown of these requests.

Request Type	N
Interruptions while Helper is talking	1
Speaking during a gap in Helper's speech	4
Preempted Helper attention changes (help needed)	6
Preempted Helper attention changes (help not needed)	1

Table 1. Breakdown of attention requests

Figure 8 shows a Worker (W1) attempting to interrupt a Helper twice while the Helper ignores the requests and continues W1 until he reaches the end of an instruction.

Figure 9 shows an attempt by a Worker (W1) to preempt the Helper's switch to the other Worker (W2).

Anticipated Presentation

Occasionally, Workers who did not have the Helper's attention would orient their pieces so that the Helper could more quickly assess the current state of the work immediately after switching attention. The Worker would then take the first turn of the joint project by asking a question such as "Is this right?" or "Like this?" We refer to this behavior as *anticipated presentation* (See Figure 10).

In the different task condition, 8 out of 144 attention switches were anticipated presentations. Except for one

session, anticipated presentation was not observed in most of the same task conditions.

H:	<i>[To W2] don't- don't worry about the gray parts anymore, we're just working at the opposite end of the purple one.</i>
W2:	<i>Do you know where this should go?</i>
H:	<i>Um.</i>
W1:	<i><interrupting> Are you sure this is right? Cause this can't grab anything</i>
H:	<i>[To W2] Yeah, that should go, on the purple end.</i>
H:	<i>No, in-between the two.</i>
H:	<i>In-between the two, all the way on the far end, by your right hand.</i>
W2:	<i>Right here?</i>
H:	<i>Even further.</i>
W2:	<i>Oh right here, I see.</i>
H:	<i>Yes.</i>
W2:	<i>OK.</i>
W1:	<i><interrupting> John,</i>
H:	<i>[To W2] And you should also be able to put one of those purple pieces in line, between that black thing that you just put in, and the purple one.</i>
	<i><Helper turns to W1></i>
W1:	<i>Like do I attach this like this but can't grab anything?</i>

Figure 8. Worker Interruption.

	<i><Helper is helping W1 but decides that W2 needs more help></i>
H:	<i>Well then you need to... <stops his own utterance></i>
	<i><Helper turns to W2></i>
W1:	<i>It attaches, but what angle should it attach at?</i>
W2:	<i>This goes...</i>
H:	<i>[to W1 while looking at W2] Give me a second</i>
W1:	<i>Oh, I got it, John.</i>

Figure 9. Floor Control by Helper.

	<i>Helper is finishing with W2.</i>
H:	<i>Exactly</i>
H:	<i>So, like that</i>
	<i>Helper turns to W1.</i>
W1:	<i>Should I be attaching these together?</i>
H:	<i>Yes,</i>

Figure 10. An example of anticipated presentation.

DISCUSSION

In this study, we examined how Helpers allocated their attention and communication across two Workers. We found that the structure of the conversation differed depending on whether the Helpers were helping both Workers on the same task or each Worker on different tasks. Based on the machine learning analysis we have done, we did not find that speech markers and other features of the interaction were predictive of the end of a joint project. We also observed several different types of attention requests and found that Workers could anticipate an attention shift by the Helper to themselves, present their work to the Helper, and take the opening turn of a joint project by asking a question. These findings have both theoretical and design implications.

Machine Learning Issues

We had hypothesized that speech markers could indicate the start and end of joint projects of attention between Helper and a Worker, as Bangerter *et al.* [2] had found. Although the entries of joint projects were likely to begin with a discourse marker, we found no regular speech markers that indicated the end of a joint project.

An error analysis reveals that we may have much more success with an alternative approach. A close examination of the conversation logs reveals that the linguistic markers we expected to be highly predictive of the ends of projects were routinely used to mark the ends of subtasks. Since several subtasks involving the same worker may occur within a single joint project, these linguistic markers are giving "mixed signals" to our trained prediction model. Furthermore, because any number of subtasks involving the same worker may occur within the same project, there was a great deal of variance in the total length of a project, which makes the amount of text or amount of time since the last project boundary less predictive than it should be of nearness to the end of a project. However, there was much less variance in the length of a subtask. Thus, we expect that if we took a similar approach, but predicted nearness to the end of a subtask rather than nearness to the end of a project, we might be able to achieve a much better result.

Another reason the markers may not have predicted joint project boundaries is due to the fact that conversations in our study were face-to-face, whereas those studied by Bangerter *et al.* were over the telephone. In our study, Helpers would initiate joint projects without speech, simply by turning their heads and reorienting their bodies. The fact that we observed anticipated presentations by Workers suggests that Helpers' body orientation may have been sufficient as a marker of the beginning of a joint project.

Since the prediction of attention was only a secondary focus, we also might have had better results had we recorded and extracted features from the Workers' workspaces. In a more recent study [18], we found that attention can be more predicted accurately (~80%) with speech and the inclusion of Workers actions in building the prediction model. In that study, we setup the experiment such that the Helper was assisting two Workers simultaneously over a video system and the Helper manually switched focus between Workers. Worker actions were recorded and included in the prediction model. That experimental setup is closer to the design scenario envisioned for this project where the Workers are actually remote and don't have access to the Helper's non-verbal cues. Also, we used a different method for preprocessing the data before constructing the machine learning models.

Theoretical implications

Preemption of Helper switches, wherein a Worker tries to avert a shift in the Helper's attention, was an interesting and unexpected phenomenon that would have been hard to

predict from past research on one-on-one assistance. We propose that preemption is evidence of breakdown in the grounding process. Workers prevent the Helper from switching to the other Worker when they realize do not understand the last instruction.

An alternate view of the Helper's act of prematurely switching away from a Worker (i.e., before full grounding is achieved) is that is an efficient adaptation to the need to multi-task. If there is no immediate reason to stay with the current Worker and provide assistance, why wait? If the Worker does come up with a question about the last instruction, then immediately switching back to that Worker is cheap in terms of grounding because the context of that Worker's task is still available in the Helper's memory. Another possible explanation is that grounding is a process that people are not conscious of and speakers only seek evidence of understanding as a natural move in continuing a conversation; if a speaker is not going to continue the conversation, they will not bother to check that their contributions have been grounded before leaving.

When Helpers returned to a Worker, they would sometimes ask for an update on the Worker's task state with questions like "What part are you on?" or "so where are you?" We think of these phrases as *regrounding* utterances, where the Helper is trying to re-establish the common ground with the Worker. Regrounding is needed for several reasons. The task environment could have changed since the Helper's last visit. The Helper also can assess the Worker's progress since the last instruction. Or, the Helper could have simply forgotten where they in that Worker's task and needs to rely on the Worker for that information.

Design implications

Our goal for this study was to observe how one expert assisted multiple novices in a face-to-face setting, in order to identify requirements for a system that would allow experts to assist multiple novices remotely. Our findings have several design implications:

- Different policies will be needed for attention shifts in situations where experts are providing instruction on the same task to all novices versus different tasks for each novice. When novices are performing the same tasks, experts must glance rapidly back and forth among them; when novices are performing different tasks, experts must provide installments of the instructions to one novice before shifting.
- Different dialogue modes may be needed for situations in which novices are working on the same versus different tasks. When on the task is the same, party-line audio may be preferable whereas a series of two-way connections may be better when the tasks are different.
- Speech markers alone cannot be used as reliable indicators of when a joint project is ending. Thus, simple solutions to managing the experts' attention, such as keyword spotting, will not work. Machine

learning algorithms did in fact perform well although further work will be required to understand how these algorithms could be implemented in a working system.

- Novices will need methods for requesting and maintaining experts' attention, and feedback about whether this feedback is being received. They will also need feedback on whether or not they are the expert's current focus of attention.
- Experts may need the system's help to mediate requests for attention and prevent disruption of their current instruction-giving efforts. Figures 8 and 9 show how, even in a situation with two Workers and face-to-face communication, requests for attention can be an issue if Workers are assertive in seeking the Helper's attention.
- Experts may gain from features that help them reground themselves when returning attention to a novice, such as a visual history of that novice's workspace since their last joint project.

Limitations and Future Directions

There are several limitations that may affect how well our results generalize to real-world expert-novice communication. First, we used a scenario in which an expert had to help only two novices at a time. It is not clear how well our findings will generalize to cases with larger numbers of novices. With two novices, each novice always knows that he or she is next in line, once the expert is finished with another novice. Greater competition for the experts' attention may arise when there are many novices.

Second, we used a simplified experimental task that, while it captures many of the essential elements of real-world collaboration on physical tasks, is still relatively easy. Thus, it may have been easier for experts to determine how well novices understood their instructions than is the case when experts provide guidance on more complex tasks.

Third, the novices in our study may not have had the same level of incentive to get the experts' attention and finish the task quickly as they might in real-world settings.

Finally, our analysis of markers for the beginning and end of joint projects did not include feedback on the novices' actions. In face-to-face settings, instructors can see if their instructions have been carried out correctly. For future studies, we plan to gather visual information on each novice's workspace.

CONCLUSION

We examined how a single expert allocated attention and communication across multiple novices in a face-to-face setting. Our goal was to identify features that an automated system could use to direct a remote expert's attention among multiple novices on an as-needed basis. We found that patterns of attention across novices differed as a function of whether or not the novices were performing the same task or different tasks. We also found that machine learning algorithms using features of the experts' speech

could reliably predict shifts between novices. In addition, the novices themselves exhibited behaviors that helped direct the expert's attention. We concluded with several design recommendations for a system to support remote experts who are assisting multiple novices.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants #0208903 and # 032977. We also thank Joe Ilorete, Pablo Quinones, Katie Waite, Mariesa Cash, Jared Metter, Julie Ludwick, and Gail Kusbit for their assistance on this project.

REFERENCES

- [1] Aoki, P. M., Romaine, M., Szymanski, M. H., Thornton, J. D., Wilson, D., & Woodruff, A.. The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. *Proc. CHI 2003*, ACM Press (2003), 425-432.
- [2] Bangerter, A., Clark, H. H., & Katz, A. R. Navigating joint projects in telephone conversations. *Discourse Processes*, 37 (2004), 1-23.
- [3] Burges, C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*. (1998)
- [4] Clark, H. H. *Using language*. Cambridge University Press, Cambridge, UK, 1996.
- [5] Clark, H. H. & Brennan, S. E. Grounding in communication. In L. B. Resnick, R. M. Levine, & S. D. Teasley (Eds.) *Perspectives on socially shared cognition*, APA (1991), 127-149.
- [6] Clark, H. H. & Wilkes-Gibbs, D. Referring as a collaborative process. *Cognition*, 22 (1986), 1-39.
- [7] Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S. & Bruce, V. Face-to-face and video mediated communication: a comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(1997), 105-125.
- [8] Endsley, M. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37 (1995), 32-64.
- [9] Fussell, S. R., Setlock, L. D., & Kraut, R. E. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. *Proc. CHI 2003*, ACM Press (2003), 513-520.
- [10] Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A. Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction*, 19 (2004), 273-309.
- [11] Gergle, D., Kraut, R. E., & Fussell, S. R. Action as language in a shared visual space. *Proc. CSCW 2004*, ACM Press (2004), 487-496.
- [12] Herring, S. C. Interactional coherence in CMC. *Proc. of the 32nd Hawaii International Conference on System Sciences*. IEEE Press (1999).
- [13] Jackson, M., Anderson, A. H., McEwan, R., & Mullin, J. Impact of Video Frame Rate on Communicative Behaviour in Two and Four Party Groups. *Proc. CSCW 2000*, ACM Press (2000), 11-20.
- [14] Jefferson, G. Side sequences. In D. Sudnow (Ed.) *Studies in social interaction*. Free Press, New York (1972), 294-338,
- [15] Kraut, R. E., Fussell, S. R., & Siegel, J. Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction*, 18 (2003), 13-49.
- [16] Kraut, R. E., Miller, M. D., & Siegel, J. Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proc. CSCW 1996*, ACM Press (1996), 57-66.
- [17] Ou, J., Oh, L. M., Fussell, S. R., Blum, T., and Yang, J. Analyzing and predicting focus of attention in remote collaborative tasks. *Proc. ICMI 2005*, ACM Press (2005), 116-123.
- [18] Ou, J., Shi, Y., Wong, J., Fussell, S. R., and Yang, J. Combining audio and video to predict helpers' focus of attention in multiparty remote collaboration on physical tasks. *Proc. ICMI 2006*, ACM Press (2006), 217-224.
- [19] Ranjan, A., Birnholtz, J. P., and Balakrishnan, R. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *Proc. CSCW 2006*, ACM Press (2006), 403-412.
- [20] Smola, A. J., Scholkopf, B. *A tutorial on support vector regression*. NeuroCOLT2 Technical Report Series – NC2-TR-198-030., 1998
- [21] Stiefelhagen, R., Yang, J., & Waibel, A. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13 (2002), 928-938.
- [22] Vertegaal, R., Weevers, I., Sohn, C., Cheung, C. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. *Proc. CHI 2003*, ACM Press (2003), 521-528.