

Speech and Sketching: An Empirical Study of Multimodal Interaction

A. Adler¹ and R. Davis¹

¹MIT CSAIL 32 Vassar St, Cambridge, MA 02139, USA

Abstract

Sketch recognition can capture the sketching component of a multimodal conversation about design, but it does not capture information conveyed in the other modalities. The informal speech that accompanies a sketch often has a considerable amount of additional information. We want to develop a digital whiteboard capable of understanding both sketching and speech, and capable of participating in a conversation similar to one that the user would have with a human design partner. We conducted a user study to help us understand what kinds of conversations users would have with a whiteboard capable of recognizing a sketch. We report results that we believe will help guide the design of an effective multimodal interface, and discuss implications for system architectures.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information interfaces and presentation]: User Interfaces. - Natural language, Graphical user interfaces, Evaluation/methodology, Input devices and strategies, Interaction styles, User-centered design, Voice I/O

1. Introduction

Sketching is widely used in the early stages of design [Dav02, UWC90]. However, the sketch alone may not tell the whole story because some information is notoriously difficult to express graphically. Sketching is often accompanied by speech that, although informal, still conveys a considerable amount of information. Interaction about the design with another person helps work out details and uncover mistakes. To illustrate the importance of speech, consider the sketch of a robot provided by one of our subjects, and compare it to the photograph of the robot (Figure 1). It's hard to make any sense of the sketch alone, but the accompanying speech identified the parts of the sketch and how they fit together.

Our goal is to make the computer a collaborative partner for early design, moving beyond a system that interprets sketches alone to a multimodal system that incorporates speech and dialogue capabilities. Although there are systems that allow the user to utter simple spoken commands to a sketch system [CJM*97, DKD05, Kai05], our long-term goal is to move beyond simple commands to create a multimodal digital whiteboard capable of a more natural conversation with the user. Instead of simple spoken commands

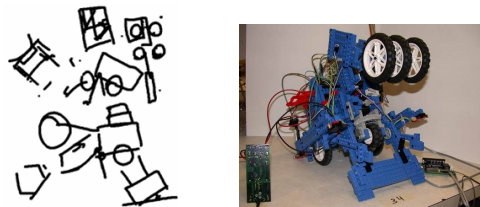


Figure 1: Left: a sketch of a robot, right: the robot.

(like uttering “block” while pointing), we want the user to provide a narrative while sketching. A complete understanding of unrestricted narrative is of course unduly difficult; our goal is to have the system understand *enough* of the sketch and *enough* of the speech to engage the user in a sensible conversation. So far we have focused on the sketching interface and the user study described in this paper. We are working on the modeling and interaction parts of the system.

Traditional dialogue and command-driven systems make many assumptions about what computer-human interaction should be like, and the dialogues are typically quite structured. Although such approaches are tractable, well-

Copyright © 2007 by the Association for Computing Machinery, Inc.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions Dept, ACM Inc., fax +1 (212) 869-0481 or e-mail permissions@acm.org.

Sketch-Based Interfaces and Modeling 2007, Riverside, CA, August 02-03, 2007.

© 2007 ACM 978-1-59593-913-5/07/0008 \$5.00

understood, and sometimes quite useful, they might not be well suited for open-ended domains such as design. To understand this issue better, we conducted a study of human-human interaction aimed at eliciting requirements for a multimodal conversational design assistant.

In the next section, we describe the user study and explain how data was annotated. The following sections describe the qualitative and quantitative results of the study, including how sketching, language, multimodal input, questions, and comments were used by the study participants. We then discuss the implications for the system architecture and conclude with a discussion of related work.

2. User Study

Other systems that let users sketch and speak are typically limited in one or more of the following dimensions:

- Command-based speech – The user talks to the system differently than they would talk to a person, issuing short commands, not natural speech. (e.g., [OD01])
- Unidirectional communication – The system cannot ask questions or add things to the sketch. (e.g., [Kai05])
- Annotation instead of drawing – The user can only annotate an existing representation, not use free form drawing. (e.g., [JEW*02])
- Fixed set of graphical symbols – The user has to know a fixed symbol vocabulary. (e.g., [OCW*00])

The goal of our study was to relax these constraints and look at a bidirectional conversation with more narrative speech and unrestricted sketching.

2.1. Study Setup

Ideally, we would have conducted a Wizard-of-Oz study in which responses to the participant would appear to be coming from the computer. Given the open-ended nature of the speech and sketching we wanted to capture, we determined that it would be too difficult to obtain a responsive and natural feeling in a Wizard-of-Oz study. We view the study as one step in developing a system which will help determine whether conversations with the computer are similar to conversations with another person.

Eighteen subjects participated in the study, all of them students in the Introductory Digital System Laboratory class at MIT. Participants were instructed to sketch and talk about four different items: a floor plan for a room with which they were familiar, the design for an AC/DC transformer, the design for a full adder, and the final project they built for their digital circuit design class. In addition, there were instructions and a warm-up condition to familiarize the participants with the system and the interface. For the AC/DC transformer and the full adder, the participants were given a textual description of the circuit and a list of suggested components. They had the option of viewing a schematic of the

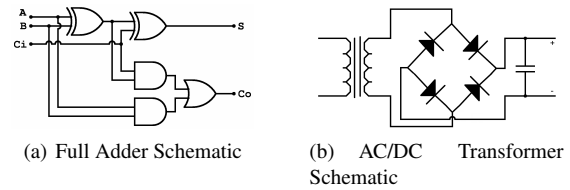


Figure 2: Schematic views of the full adder and the AC/DC transformer that the participants could choose to view.

transformer and adder circuits (Figure 2) before they began drawing, but the schematic was not visible while they were drawing.

The experimenter and participant sat across a table from each other, each with a Tablet PC. We considered having a physical barrier between the experimenter and the participant, but didn't because a barrier would have created an unnatural environment and obstructed the video recording. In order to encourage all communication to be done by interacting with the drawing surface, the experimenter looked at his tablet and avoided eye contact with the participant. The Tablet PCs were equipped with software we designed that replicates on each tablet in real time whatever is drawn on the other tablet, in effect producing a single drawing surface usable by two people at once. It is possible that interactions with the Tablet PCs differ from the interactions with a whiteboard sized device, but for this study we used Tablet PCs.

The software allowed the users to sketch and annotate the sketch using a pen and a highlighter. Buttons above the sketching area allowed users to switch between five pen colors and five highlighter colors. Another button allowed users to switch into or out of a pixel-based erase mode, allowing either user to erase parts of any stroke. Finally, there was a button that allowed either user to create a new blank page. The software recorded the (x, y) position, time, and pressure data for each point drawn by either user. To enhance the feeling of naturalness, strokes were rendered so that they were thicker when the user applied more pressure. Two video cameras and headset microphones were used to record the study. The audio, video, and sketching inputs were synchronized. This enabled playback and analysis of the timing of the speech and sketching events. At various points in the study, the experimenter added to the sketch and asked questions about different components.

2.2. Data Annotation

At the conclusion of the study, we had two movie files (one for the participant and one for the experimenter) for each of the four items the users were asked to draw, along with one XML file [Ske] for each page of sketching. The XML files contained a full record of the sketching by both the participant and the experimenter.

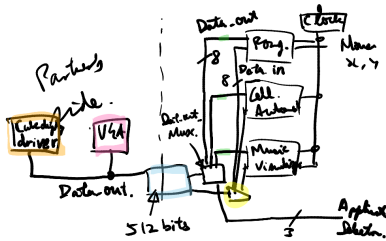


Figure 3: A sketch of a participant's project.

We created software that replayed the study by using these data streams. The software also allowed us to select parts of the audio tracks for playback and transcription. The transcript and audio segments were passed to the Sphinx speech recognizer [HAH*93] forced-alignment function, which produced precise timestamps for each word. The transcripts were verified by playing the segment of the audio file and confirming that it contained the correct word.

3. Study Analysis

Our analysis of the study has focused on how speech and sketching work together when people are interacting with each other. Figure 3 shows a sketch, and Figure 4 illustrates the type of speech that accompanied it. In general the sketches contained the circuit itself and additional strokes related to its function or identification of its components. In Figure 5 the sketch contains the AC/DC converter and arrows indicating the flow of current in each of two operating conditions. Highlighter strokes are used to identify components in the circuit.

Data from 6 of the 18 participants have been processed as described above. Each of the six datasets has data from each task (i.e., the warm-up and four sketching tasks). The total length of the data is approximately 105 minutes; about 17.5 minutes of data for each participant. The participants drew 2704 strokes, 74 erase strokes, and spoke 10,848 instances of 1177 words. The experimenter drew 155 strokes, 3 erase strokes, and uttered 2282 instances of 334 words. The rest of the data has not yet been transcribed because of the time-consuming nature of the transcription and annotation process.

Our ongoing qualitative analysis of the recorded and transcribed data has led to a series of initial observations. We have divided the observations into five categories: sketching, language, multimodal interaction, questions, and comments. Although these categories aren't mutually exclusive, they help frame the observations and our discussion.

3.1. Observations about Sketching

Our observations about sketching can be divided into two categories: stroke statistics and the use of color.

Experimenter: so then what's what's um this piece what's that
 Participant: that would be the mux for the data input actually

Participant: that was a uh uh yeah a memory bank with five hundred and twelve um yep five hundred and twelve bits this ah i could that i had read and write access to

Figure 4: Fragments of the conversation accompanying Figure 3. Notice the disfluencies and repeated words.



Figure 5: A sketch from the user study of an AC/DC transformer.

As part of our analysis, we labeled each stroke as one of four types: creation, modification, selection, and writing. Creation strokes accounted for 52% of the strokes, writing strokes accounted for 40%, selection strokes accounted for 5%, and modification strokes accounted for 3%. Looking at the total amount of ink in the sketches, 63% was from creation strokes, 21% was from writing strokes, 8% was from selection strokes, and 8% was from modification strokes. The percentage of ink that is writing more accurately reflects the composition of the sketches than stroke count because multiple strokes are frequently used for a single letter. A low percentage of strokes are selections, but these are very important strokes to understand because they are key to understanding the user's action.

The average number of colors used in a sketch was 3.0 (with a standard deviation of 1.8). The number of times the color was changed in a sketch was much more variable; the average number of color changes was 4.7, but the standard deviation was 6.5. There were a few sketches where the participant changed the color more than 10 times; once during a long sketch, a participant changed the color 28 times.

We found that color was used in several different ways: to identify regions that were already drawn, to differentiate objects, and to add an "artistic" character. We consider each of these in turn.

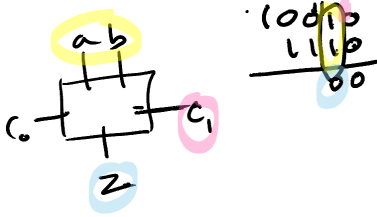


Figure 6: Color was used to indicate corresponding areas of the sketch.

Color was frequently used to refer back to existing parts of the sketch and/or to link different parts together. In Figure 6, for example, three different colors were used to indicate the correspondence between different parts of the sketch, a crucial step in understanding it.

We found that a color change is an excellent indication that the user is starting a new object. Segmenting strokes into objects is difficult because there are numerous ways to group the strokes. Color can aid segmentation by providing a good clue about which strokes should be grouped together.

Color was also used to reflect the real-world colors of objects, including for example, bodies of water. These colors can aid in segmenting the input, but also have deeper meaning because they relate to real-world objects and associations. This would allow color references in the speech to be matched with the colors and objects in the sketch.

Others have explored the use of color in sketching, e.g., Classroom Presenter [AHWA04]. They found that the frequency of color changes varied by presenter, but that color changes were used for contrast and to visually distinguish objects. Our more complicated sketches seemed to use more colors which concurs with the findings of [AHWA04].

3.2. Language

The language chosen by participants provided several valuable insights. The most readily apparent observation is that the speech tended to be highly disfluent, with frequent word and phrase repetition. This phenomenon appears to occur more frequently when participants are thinking about what to say. Second, participants' responses to questions posed to them tended to reuse words from the question. Third, not unexpectedly, the speech utterances are related to the current sketching. We address each of these observations in turn.

The repetition of words or phrases occurred more frequently when participants were thinking about what to say. One participant, describing the output "R" of a circuit, said: "the result will be R, whereas... if so let's let's eh the result will be R... is that if the carry in is carry eh if the carry in is one, then the result here will be R, this is in case the carry in is one." The speech here is ungrammatical, disfluent,

and repetitive, clearly making it more difficult for a speech recognition system. However, the repetition of the key words "result," "carry in," and "R" should allow us to identify them as the key concepts being discussed. The repetition could also provide evidence that the user is thinking about what to say. This evidence about user uncertainty could help a system determine that the user is interruptible.

Participants' responses to questions tended to reuse vocabulary from the question. For example, when asked "so is this the, is that the diode?," the participant replied: "this is the diode, yeah." A system could expect a response to questions to have phrasing similar to the question, facilitating the speech recognition task.

Not unexpectedly, we found that the participants' speech relates to what they are currently sketching. For example, in one sketch the participant is drawing a box and while drawing it says "so let's see, we got the power converter over here;" the box is the representation of the power converter he is talking about. This may facilitate matching the sketching and speech events as they are roughly concurrent.

3.3. Multimodal

We encountered three varieties of multimodal interaction between the speech and sketching inputs: referencing lists of items, referencing written words, and coordination between input modalities.

Participants in the study would often verbally list several objects and sketch the same objects in the same order. For example, when sketching a floor plan, one participant said "eh so here I got a computer desk, here I got another desk, and here I got my sink," while sketching the objects in the same order. In another sketch, a participant drew a data table and spoke the column labels aloud in the same order that he sketched them. The consistent ordering of objects in both modalities provides another method for associating sketched objects with the corresponding speech.

Participants who wrote out words such as "codec" or "FPGA" referenced these words in their speech, using phrases such as "so the the codec is pretty much built in, into the, like uh standard, um, eh, standard, uh FPGA interface." If the handwriting can be recognized, this information can help identify the words in the speech input, as has been done in [Kai05]. Participants also wrote abbreviations for spoken words, for example, "Cell." for "Cellular." Recognizing these textual abbreviations will also help find correspondences between the sketch and the speech.

As noted, we found that speech often roughly matches whatever is currently being sketched. Subjects indicated a tendency to enforce this coordination: if a subject's speech got too far ahead of their sketching, they typically slowed down or paused their speech to compensate.

There were many examples in the study where the participant paused their speech to finish drawing an object, then

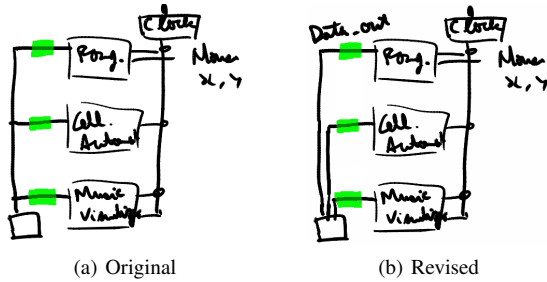


Figure 7: Left: the original sketch, right: after revision. One data output line in the original image has been replaced by three in the revised image.

continued. For example, one participant said “and that’s also a data out line” and then finished writing “Data out” before continuing the speech. In another case, a participant said “um, you come in and” and then paused while he finished drawing an arrow to indicate the entrance to the room. These observations provide additional data that the two modalities are closely coordinated. We can use this relationship in a system to help match speech utterances with sketching.

3.4. Questions

When the experimenter asked the participants questions, the participants often made revisions or explained their design in more depth than the questions required.

Some questions caused the participant to make the sketch more accurate, as in Figure 7. When the experimenter asked if the three outputs, highlighted in green (Figure 7(a)), were the same, the participant realized that the original sketch was inaccurate, then revised it by replacing one data output with three separate lines (Figure 7(b)).

Questions about one part of the sketch also spurred explanations about other parts, as participants apparently decided other parts of the sketch might be confusing as well, based on the question asked. When one participant was asked about a label for a column in a data table, he not only clarified that label, he also explained the other four labels.

Comparison questions also encouraged participants to explain the sketch in more detail by explaining how the parts were or were not similar. For example, participants were asked if several different gates in the full adder were the same. A participant’s reply might be that both gates are AND gates or that one is an AND gate and one is an OR gate.

These elaborated answers to questions were an unexpected result of the study. Asking questions keeps the participant engaged and encourages them to continue talking. The resulting additional speech and sketching data would give a system a better chance of understanding the sketch. The interaction also appears to encourage the participants to

provide more information about the sketch, and it appears to cause the participants to think more critically about the sketch, so that they spot and correct errors or ambiguities. Even simple questions like “Are these ___ the same?” seems to be enough to spark an extended response from the participant, especially if there is a subtle aspect of the objects that was not previously revealed.

3.5. Participant Comments

Participants made several comments during the study that did not relate directly to the sketch, but still provided valuable information. Uncertainty was indicated through the use of phrases such as “I believe” or “I don’t remember.” Some comments related to the user interface, for example, “I’ll try to use a different color.” Other comments referenced the appearance of the sketch. Two examples of this type of comment are: “it’s all getting a little messy” and “I’ll draw openings like this I don’t know... I draw li... I drew like a switch before.” These comments still provide insight into the participant’s actions and could help a system understand what the users are doing, but don’t relate directly to the sketching.

Another observation from the study is that the participant and the experimenter are expected to be able to fill in words that their partner forgot. For example, one participant expected the experimenter to help with forgotten vocabulary, and another participant filled in a word the experimenter forgot. This might be another way that a system could interact with the user, saying something like “And this is ah...” and pausing, prompting the user to identify the object.

4. Quantitative Analysis

Work in [ODK97] reports on a series of user studies in which users interacted multimodally with a simulated map system. They examined the types of overlap that occurred between the speech and sketching, finding that the sketch input preceded the speech input a large percentage of the time. The nature of the overlap is important to properly align the speech and sketching inputs. More recently, [KBEC07] examined the timing of speech and handwriting.

We performed a similar analysis of our data, matching corresponding speech phrases and sketching events. For example, we matched the speech utterance “so we have an arrangement of four diodes” with the strokes making up the diodes that were sketched concurrently. We segmented the speech into phrases based on pauses in the participants’ speech. We call these *phrase groups*.

Within each phrase group, we created groups containing only a word and the strokes it was referring to, for example, the word “diode” and the strokes making up the diode. We’ll call these *word groups*. These two types of groups were generated in light of differences in the nature of overlap between the speech and the sketching events as compared to



















Speech Precedes (82%)	Sketch Precedes (16%)	Neither Precedes (2%)
 Sketch  Speech (1%)	 Sketch  Speech (1%)	 Sketch  Speech (0%)
 Speech  Sketch (30%)	 Speech  Sketch (5%)	 Speech  Sketch (0%)
 Speech  Sketch (51%)	 Speech  Sketch (11%)	 Speech  Sketch (2%)

Table 1: The temporal overlap patterns for the phrase groups. The alignment of the speech and sketching is shown in an illustration in each table cell, along with the percentage of phrase groups in each category.



















Speech Precedes (26%)	Sketch Precedes (71%)	Neither Precedes (3%)
 Sketch  Speech (0%)	 Sketch  Speech (1%)	 Sketch  Speech (0%)
 Speech  Sketch (24%)	 Speech  Sketch (14%)	 Sketch  Sketch (3%)
 Speech  Sketch (2%)	 Speech  Sketch (55%)	 Sketch  Sketch (0%)

Table 2: The temporal overlap patterns for the word groups.

the results from [ODK97]: [ODK97] said sketching usually precedes speech, which matches the results from our *word groups*, however the results of our *phrase groups* indicate that speech usually precedes sketching.

We compared the start time of sketching with the start time of speech, and compared the end time of sketching with the end time of speech. Table 1 shows the nine possible ways the speech and sketching can overlap and the percentage of time each occurred for the phrase groups. Table 2 shows the same thing for the word groups. These are the same temporal overlap patterns used in [ODK97].

Unlike the videotape analysis used in [ODK97], we have precise timing data for our speech and pen input, both measured in milliseconds. By analyzing the video of several speech/sketching groups whose overlap difference was very small, we determined that time differences of less than 50 milliseconds were undetectable, hence we label events as simultaneous if they are less than 50 milliseconds apart. The video was recorded at 30 frames per second (approximately one frame every 33 milliseconds).

The histograms in Figure 8 and Figure 9 reflect the differences between the start times for the sketching and the speech events in each group. The x-axis is the time in milliseconds that the start of the sketching preceded the start of the speech (a negative number means that the speech pre-

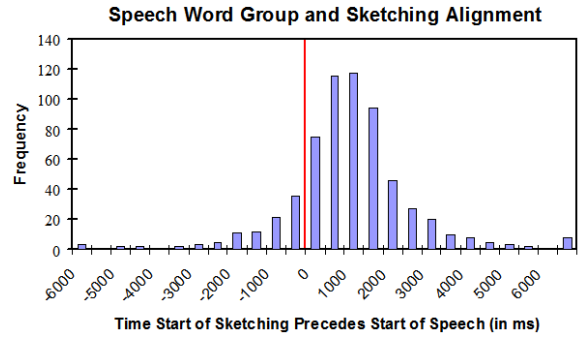


Figure 8: A histogram depicting the time difference between the start time of the speech and sketching in each word group. The x-axis is the time in milliseconds that the start of the sketching preceded the start of the speech.

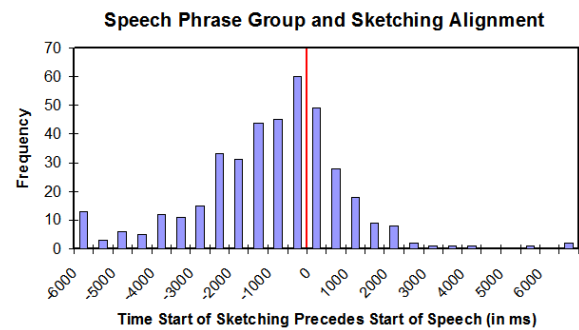


Figure 9: A histogram depicting the time differences between the start time of the speech and sketching in each phrase group.

ceded the sketching). We have the data for both differences in start and end times, but we chose this visualization because it emphasizes the difference between the word level and phrase level groupings.

Figure 8, depicting time differences for the word groups, shows that in most cases (72%), the sketching precedes the spoken word; these data points are in the right half of the histogram. The data is also tightly clustered near zero; this shows that sketching occurred temporally near the speech that referenced it.

The results for the *word groups* match the results reported by [ODK97]. They reported that 57% of the time writing preceded speech (our data shows 72%). The most frequent overlap category they had was sketching starting first and ending first; this was also our highest category for the word groups (55%). We also examined only the word groups that contained only handwriting to compare the results to [KBEC07]. The results are very similar to [KBEC07], with

handwriting preceding speech 63% of the time and speech preceding handwriting 35% of the time.

We also examined the overlap that occurred in the *phrase groups*, as shown in Table 1 and Figure 9. The phrase plot shows a different relationship from the word plot. Most of the data points are in the left half (82%) representing phrases where the speech preceded the sketching.

This is the opposite of the data reported in [ODK97], which reported that sketching usually preceded the speech. There are several possible explanations for this difference. Their study looked at users sketching on an existing map, and our study examined users drawing on a blank page. Our users explained the function of the various parts of their designs – something that doesn't happen when locating places on a map. In our task the users provided narrative speech, sometimes lasting 20 seconds, when describing the sketch; this differs from the command-type speech in [ODK97]. The outliers in our histograms are the results of these long sections of speech; e.g., a few seconds of sketching at the end of a long section of speech produces a large difference between the starting times of the speech and sketching.

The difference may also arise because [ODK97] used a Wizard-of-Oz study, so the participants were talking to a computer instead of a person across the table. Our person-to-person interactive conversation could also have had an effect on the timing and type of speech and sketching data that was observed.

The mean of the difference between start times of the speech and sketching events in Figures 8 and 9 is statistically different from zero. The phrase data mean is -1464 ms and is significant ($t(397) = -13.2, p < .01$); the word data mean is 670 ms and is significant ($t(628) = 9.60, p < .01$). We also ran a two-sample t-test assuming unequal variances and the phrase data differences and the word data differences are significantly different ($t(705) = -16.3, p < .01$). Our analysis of the time differences between the onset of speech and the onset of sketching provides some data on how the system should expect the speech and sketching inputs to relate to each other.

5. Architectural Implications

Our analysis of the study provides some implications for the architecture of a multimodal digital whiteboard. Such a system should integrate knowledge about color and recognize the importance of switching ink colors. Speech and sketching are closely connected through timing and object references. An early integration of the data sources would allow a system to capitalize on this relationship. For example, lists occur in the same order in both modalities and an early integration can take advantage of this pattern. Although the repetitive and disfluent nature of the speech may make it difficult to parse entire sentences correctly, recognizing repeated

words would be easier and appear to be advantageous. Asking the user questions is important to clarify information and to encourage a rich dialogue with the user. Using the observations from the study to guide our architectural choices will strengthen the resulting system.

6. Related Work

Wizard-of-Oz studies are common and have been conducted even in situations where the wizard simulates both pen and speech data [OCFF92, ODK97]. However, in those studies the pen input is not open ended and the wizard can have a good idea of what the user will draw. In our case, we sought to remove as many restrictions as possible to gather data indicative of a conversation between two people. All of the participants in our study had unique design projects that they described, which were unknown before the user study.

QuickSet [OCW*00] is a collaborative multimodal interface that recognizes sketched icons. The user can create and position items on a map using voice and pen-based gestures. For example, a user could say “medical company facing this way <draws arrow>.” QuickSet is command-based, targeted toward improving efficiency in a military environment. It also differs from our desired system because the users have a map to refer to instead of a blank screen to sketch on.

Focusing explicitly on managing multimodal dialogues, Johnston et al. describe MATCH in [JEW*02]. MATCH includes a finite state transducer based component for combining multimodal inputs, including speech, sketch, and handwriting, in the domain of map-based information retrieval. MATCH's dialogue manager enables a goal-directed conversation, using a speech-act dialogue model similar to [RS98]. This tool provides some multimodal dialogue capabilities, but it is not a sketching system and has only text recognition and basic circling and pointing gestures for the graphical input modality.

Several existing systems allow users to make simple spoken commands to the system [DKD05, Kai05]. We had many instances of users writing words and speaking them, which is very similar to the types of input that [Kai05] handles. Kaiser describes how they can add new vocabulary to the system based on handwritten words and their spoken equivalents of the type that appear in Gantt schedule-charts [Kai06].

ASSISTANCE [OD01] was a previous effort in our group to combine speech and sketching. It built on ASSIST [AD01] by letting the user describe the behavior of the mechanical device with additional sketching and voice input. More recently we built a system [AD04] that let users simultaneously talk in an unconstrained manner and sketch. This system had a limited vocabulary and could not engage the user in a dialogue, limiting its ability to interpret the user's input.

7. Conclusion

We want to develop a digital whiteboard that can understand and participate in a natural conversation with a user who is engaged in a design task. We conducted a study to gather data about natural conversations about designs and to help guide the design of such a system. The focus of the study was to examine how speech and sketching work together when people interact with each other. The data from the user study provided some initial qualitative observations about sketching, language, multimodal interactions, questions, and comments. We found that speech phrases preceded sketching, contrary to earlier studies, but the ordering of individual words and the corresponding sketch element matched the results from earlier research. The data from the study will help us to achieve our long-term goal of moving beyond simple commands to create a multimodal system in which the user can have a more natural conversation with the computer about design.

8. Acknowledgments

This work is funded by the MIT Oxygen Project and by Pfizer, Inc. The authors would also like to thank all the participants in the user study.

References

- [AD01] ALVARADO C., DAVIS R.: Resolving ambiguities to create a natural sketch based interface. In *IJCAI* (2001). 7
- [AD04] ADLER A., DAVIS R.: Speech and sketching for multimodal design. In *IUI* (2004), ACM Press, pp. 214–216. 8
- [AHWA04] ANDERSON R. J., HOYER C., WOLFMAN S. A., ANDERSON R.: A study of digital ink in lecture presentation. In *CHI '04* (New York, NY, USA, 2004), ACM Press, pp. 567–574. 4
- [CJM*97] COHEN P. R., JOHNSTON M., MCGEE D. R., OVIATT S. L., PITTMAN J., SMITH I., CHEN L., CLOWI J.: Quickset: Multimodal interaction for distributed applications. In *Multimedia* (1997), ACM Press, pp. 31–40. 1
- [Dav02] DAVIS R.: Sketch understanding in design: Overview of work at the MIT AI lab. *AAAI Spring Symposium* (2002), 24–31. 1
- [DKD05] DEMIRDJIAN D., KO T., DARRELL T.: Untethered gesture acquisition and recognition for virtual world manipulation. *Virtual Reality* 8, 4 (2005), 222–230. 1, 7
- [HAH*93] HUANG X., ALLEVA F., HON H.-W., HWANG M.-Y., ROSENFELD R.: The SPHINX-II speech recognition system: an overview. *Computer Speech and Language* 7, 2 (1993), 137–148. 3
- [JEW*02] JOHNSTON M., EHLEN P., WALKER M., WHITAKER S., MALOOR P.: MATCH: An architecture for multimodal dialogue systems. In *ACL* (2002), pp. 276–383. 2, 7
- [Kai05] KAISER E. C.: Multimodal new vocabulary recognition through speech and handwriting in a whiteboard scheduling application. In *IUI '05* (New York, NY, USA, January 2005), ACM Press, pp. 51–58. 1, 2, 4, 7
- [Kai06] KAISER E. C.: Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *ICMI* (Banff, Canada, 2006). 7
- [KBEC07] KAISER E. C., BARTHELMESS P., ERDMANN C., COHEN P.: Multimodal redundancy across handwriting and speech during computer mediated human-human interactions. In *CHI '07* (New York, NY, USA, April 2007), ACM Press, pp. 1009–1018. 5, 7
- [OCFF92] OVIATT S., COHEN P., FONG M., FRANK M.: A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *ICSLP* (1992), University of Alberta, pp. 1351–1354. 7
- [OCW*00] OVIATT S., COHEN P., WU L., VERGO J., DUNCAN L., SUHM B., BERS J., HOLZMAN T., WINOGRAD T., LANDAY J., LARSON J., FERRO D.: Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. In *HCI* (2000), pp. 263–322. 2, 7
- [OD01] OLTMANS M., DAVIS R.: Naturally conveyed explanations of device behavior. In *PUI* (2001). 2, 7
- [ODK97] OVIATT S., DEANGELI A., KUHN K.: Integration and synchronization of input modes during multimodal human-computer interaction. In *Human Factors in Computing Systems* (1997), ACM Press, pp. 415–422. 5, 6, 7
- [RS98] RICH C., SIDNER C.: COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapter Interaction* 8, 3–4 (1998), 315–350. 7
- [Ske] MIT SketchML Format. <http://rationale.csail.mit.edu/ETCHASKETCHES/format/>. 3
- [UWC90] ULLMAN D. G., WOOD S., CRAIG D.: The importance of drawing in the mechanical design process. *Computers and Graphics* 14, 2 (1990), 263–274. 1