Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

Proposal for Thesis Research in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Title:	Multimodal Interactive Digital Whiteboard
Submitted by:	Aaron Adler 32 Vassar St., 32-239 Cambridge, MA 02139
	________________________ (Signature of Author)
Date of Submission:	February 2, 2007
Expected Date of Completion:	May 2008
Laboratory:	Computer Science and Artificial Intelligence Laboratory

Brief Statement of the Problem:

Sketching is used in early stage design, but it doesn't capture or provide the user all the avenues of communication that talking to a design partner does. By creating an interface that combines speech and sketching, we hope to engage the user more fully and capture and understand more of the sketch in a manner similar to a human design partner.

We envision a multimodal interactive dialogue. As the user talks and sketches, the computer attempts to understand. It can ask questions to resolve uncertainties, inquire about vocabulary, or ask about similar components in the sketch. These questions could be asked using a combination of verbal and sketched output.

We have conducted user studies to gather data about how two people naturally interact during design discussions, and will use the results to help guide the design of our system.

Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science
Cambridge, Massachusetts 02139

Doctoral Thesis Supervision Agreement

TO: Department Graduate Committee
FROM: Professor Randall Davis

The program outlined in the proposal:

TITLE: Multimodal Interactive Digital Whiteboard
AUTHOR: Aaron Adler
DATE: February 2, 2007

is adequate for a Doctoral thesis. I believe that appropriate readers for this thesis would be:

READER 1: Professor Robert Miller
READER 2: Doctor James Glass

Facilities and support for the research outlined in the proposal are available. I am willing to supervise the thesis and evaluate the thesis report.

SIGNED: ______________________
PROFESSOR OF ELECTRICAL
ENGINEERING AND COMPUTER SCIENCE

DATE: ______________________

Comments:

____

____

____

Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science
Cambridge, Massachusetts 02139

Doctoral Thesis Reader Agreement

To: Department Graduate Committee
From: Professor Robert Miller

The program outlined in the proposal:

Title: Multimodal Interactive Digital Whiteboard
Author: Aaron Adler
Date: February 2, 2007
Supervisor: Professor Randall Davis
Other Reader: Doctor James Glass

is adequate for a Doctoral thesis. I am willing to aid in guiding the research and in evaluating the thesis report as a reader.

Signed: ______________________________
Associate Professor of Electrical
Engineering and Computer Science

Date: ______________________________

Comments:

____

____

____

____

____

____

Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science
Cambridge, Massachusetts 02139

Doctoral Thesis Reader Agreement

To:	Department Graduate Committee
From:	Doctor James Glass

The program outlined in the proposal:

Title:	Multimodal Interactive Digital Whiteboard
Author:	Aaron Adler
Date:	February 2, 2007
Supervisor:	Professor Randall Davis
Other Reader:	Professor Robert Miller

is adequate for a Doctoral thesis. I am willing to aid in guiding the research and in evaluating the thesis report as a reader.

Signed: ______________________________
Principle Research Scientist

Date: ______________________________

Comments:

____
____
____
____
____
____

Contents

Chapter 1

Introduction

Sketching is widely used in the early stages of design to think about, communicate, and record designs for devices such as electrical circuits or mechanical systems [8, 25]. However the sketch alone may not tell the whole story, because some information is notoriously difficult to express using sketching alone. Sketching is often accompanied by speech that, while informal, still conveys a considerable amount of information. People often use these sketches in a discussion about the design in which participants can all sketch, speak, and ask or answer questions about the design. The interaction about the design with another person helps work out details and uncover mistakes.

Two simple examples help to illustrate the benefits of both speech and a two way dialogue for understanding design sketches. Consider the sketch of a robot provided by a subject in one of our user studies (Figure 1-1), and compare it to the photograph of the robot (Figure 1-2). It's impossible to make any sense of the sketch alone, yet it makes considerable sense when viewed along with the speech that went with it. Sketches are incredibly useful, but by themselves they can be insufficient. Information can sometimes be expressed more easily by sketching, like information about location and connectedness, and sometimes information is expressed more easily using speech, like information about properties of objects. By combining the two modalities the user is free to use which ever one is more natural.

As another example consider Newton's Cradle (see Figure 1-3), a system of pendulums that consists of a row of metal balls on strings. When you pull back a number of balls on one end, after a nearly elastic collision the same number of balls will move outward from the other end of the system. Although this system seems simple enough to sketch, it is in fact nearly impossible to draw so that it operates properly. The system works because the metal balls just touch each other, and because each pendulum is identical to the others. In the sketching system, you would have to draw identical pendulums, and align them perfectly. If the user could simply say that "there are five identical, evenly spaced and touching pendulums," the device would be far easier to create.

The benefit of speech is clear, but what happens if the sketch and speech recognition produce inconsistent results? If the number of pendulums drawn and referenced in the speech differ, without clarification the system is unable to take action. With a clarifying dialogue, the system can resolve the discrepancy and take the action the

Figure 1-1: A sketch of a robot.

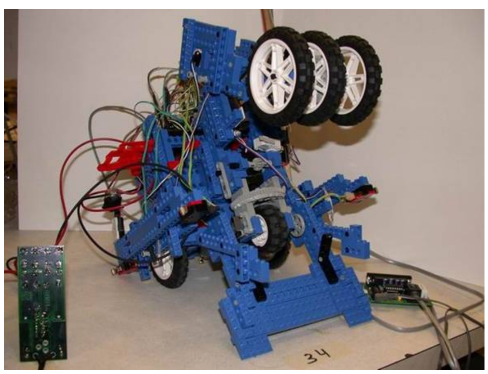


Figure 1-2: The robot sketched in Figure 1-1.

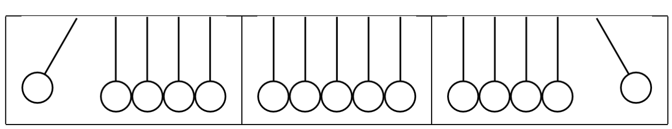


Figure 1-3: A sequence of images showing Newton's Cradle when one of the pendulums is pulled back and released.

user intended.

Our goal is to make the computer a collaborative partner for early design, moving beyond a sketching system to a multimodal system that incorporates speech and dialogue capabilities. While there are systems that allow the user to utter simple spoken commands to a sketch system [7, 9, 17], our long-term goal is to move beyond simple commands to create a multimodal digital whiteboard that allows the user to have a more natural conversation with the computer. We do not want the speech to be limited to simple commands (like uttering "block" while pointing), but want instead to allow the user to say whatever comes to mind and have the system understand *enough* of the sketch and *enough* of the speech to engage the user in a sensible conversation [1].

Traditional dialogue and command-driven systems make many assumptions about what computer-human interaction should be like, and typically involve quite structured dialogues. While such approaches are tractable, well-understood, and sometimes quite useful, they might not be the optimal form of multimodal interaction 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.

The challenges faced in building such a design assistant include integrating the speech and sketching inputs, interpreting the input, and determining how certain the system is of the interpretation. The system must understand and manage the dialogue, determine what questions to ask the user, determine how and when to ask these questions, and understand the responses.

We plan to start by working on combining the speech and sketching inputs. This will enable us to then work on the dialogue manager and more advanced interactions with the user. Key contributions of this work will include studies into how humans naturally interact when discussing these early designs, and constructing a system that can carry on a multimodal dialogue with the user.

1.1 Scenario

The following scenario illustrates the kind of interactions we would like the system to have with the user. In this scenario, the user is describing a circuit that indicates whether or not a fuse is blown, using a green and a red L.E.D. (Light Emitting Diode). In these fragments, "U" represents the user's actions and "S" represents the system's actions.

1.1.1 Example 1

This example shows how the system can query the user about a component of the sketch whose symbol is unknown. In this case, the system knows the word "fuse," but does not know what the symbol for a fuse is. In order to learn the symbol for the fuse, the system asks the user to locate it in the sketch. This fragment also illustrates one way the user can identify part of the sketch – by circling the component.

U:	**Speech:** I need a circuit that will indicate that a fuse is blown. We have a DC current and it goes through this fuse here. **Sketch:** Draws DC in and DC out and then draws the fuse
S:	**Speech:** I don't know what the symbol for a fuse is. Can you show me?
U:	**Speech:** It's here. **Sketch:** Circles fuse in pink

Table 1.1: User draws the fuse.

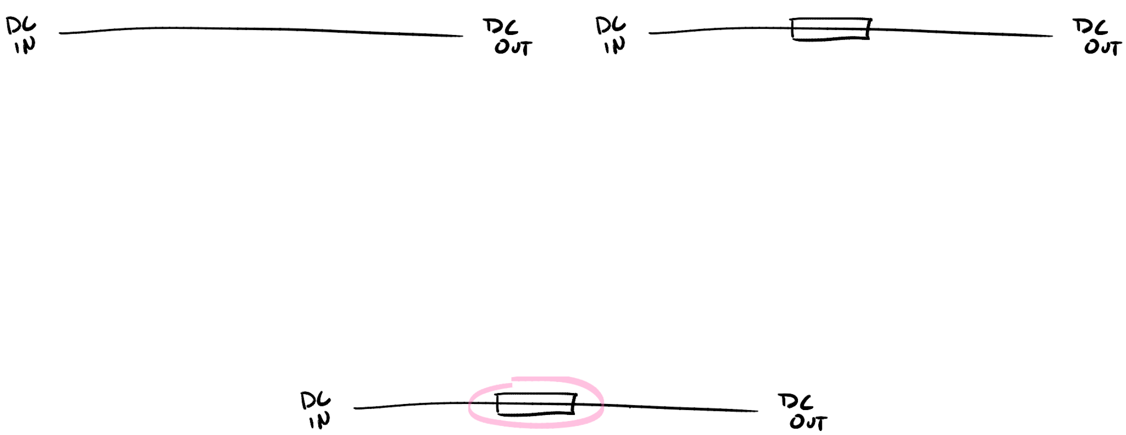


Figure 1-4: User drawing the fuse in Table 1.1.

1.1.2 Example 2

U:	**Speech:** We have a green L.E.D. here. Umm... and there is a resistor... umm... that is connec... connected to the ground. **Sketch:** Draws L.E.D., resistor, and ground
S:	**Speech:** Ok.

Table 1.2: User draws several circuit components.

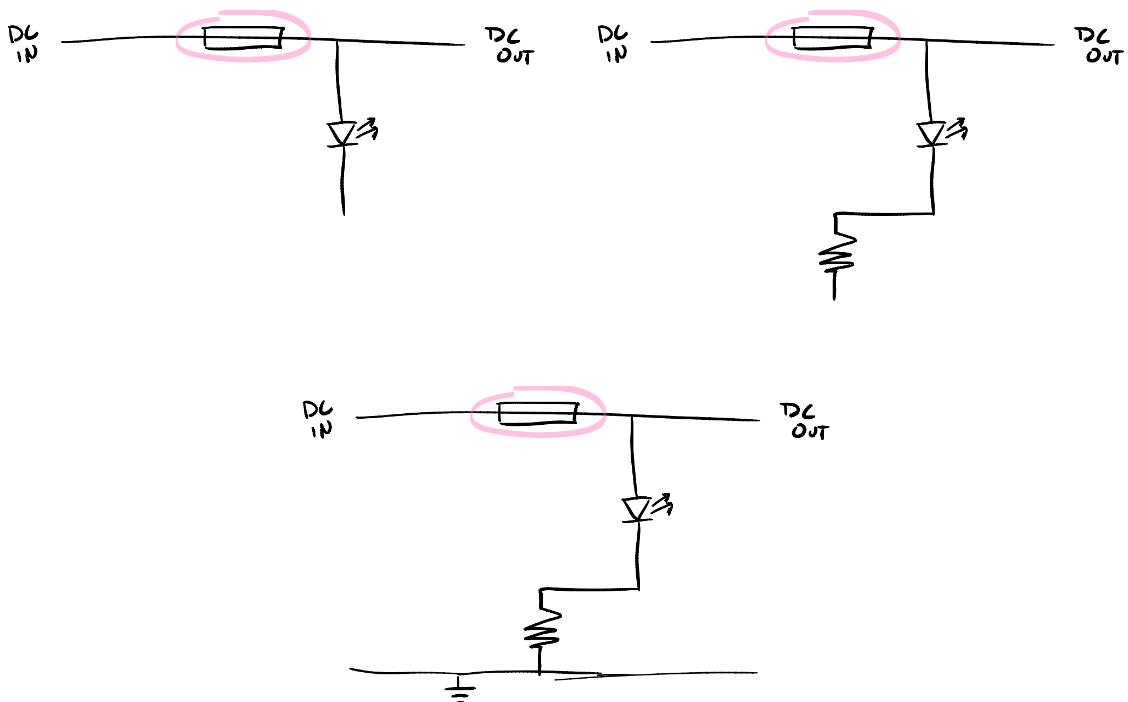


Figure 1-5: User drawing the components in Table 1.2.

In this fragment, the user draws several components and the system acknowledges this with an “ok” to indicate that the user should proceed with the explanation. The simple response is important to keep the user engaged and to make the conversation more closely simulate a conversation with another person. This fragment also illustrates disfluencies in the users speech including partial words, “umm,” and “ahh.”

1.1.3 Example 3

U:	**Speech:** On this side of the fuse we have two diodes. **Sketch:** Draws two diodes
S:	**Speech:** I found one of the diodes. **Sketch:** Circles first diode in green **Speech:** Where is the second one?
U:	**Speech:** Connected to it. **Sketch:** Circles second diode in green
S:	**Speech:** Ok.

Table 1.3: The user draws two diodes and then helps the system locate one of them.

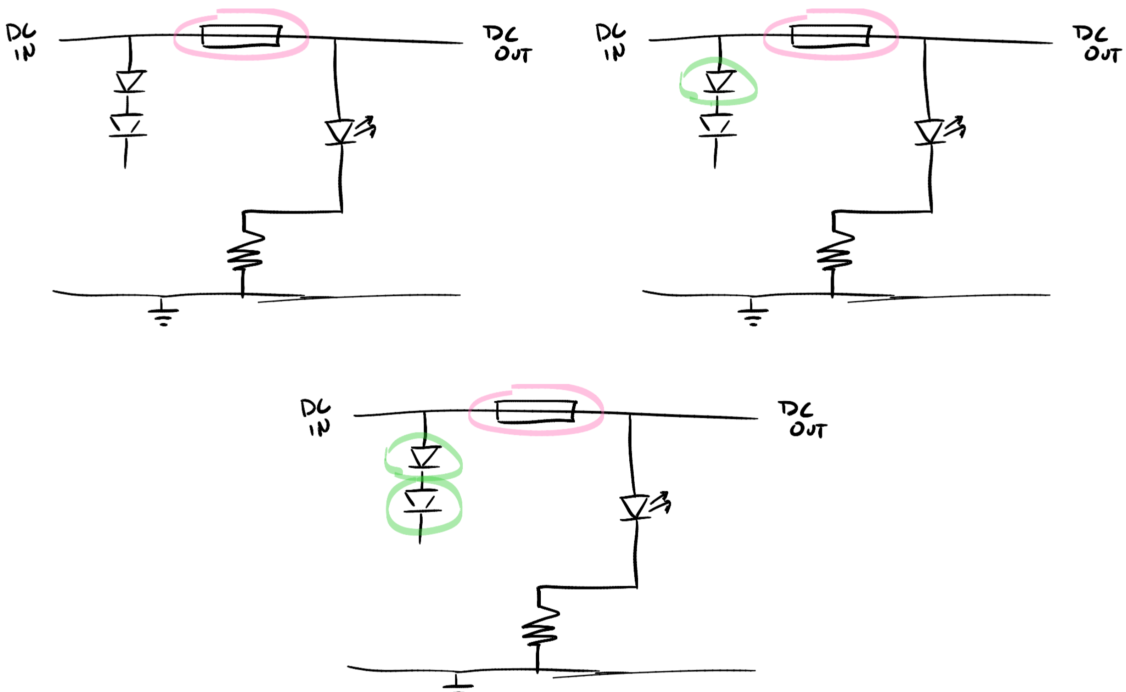


Figure 1-6: User drawing and identifying the diodes in Table 1.3.

In this fragment, the system recognizes that the user indicated that there should be two diodes, but the system only locates one of them. It then asks the user to identify the second diode. The user's answer to the question uses the same color highlighter, a phenomenon that occurred frequently in our studies.

1.1.4 Example 4

U:	**Speech:** Then we have a red L.E.D. and it is also connected to the resistor. **Sketch:** Draws the L.E.D.
S:	**Speech:** Are these the same? **Sketch:** Circles both L.E.D.s in blue
U:	**Speech:** They are similar, but, but this, this one is red... **Sketch:** Draws over black arrows of the red L.E.D. in red **Speech:** and this one is green. **Sketch:** Draws over black arrows of the green L.E.D. in green

Table 1.4: User draws the red L.E.D.

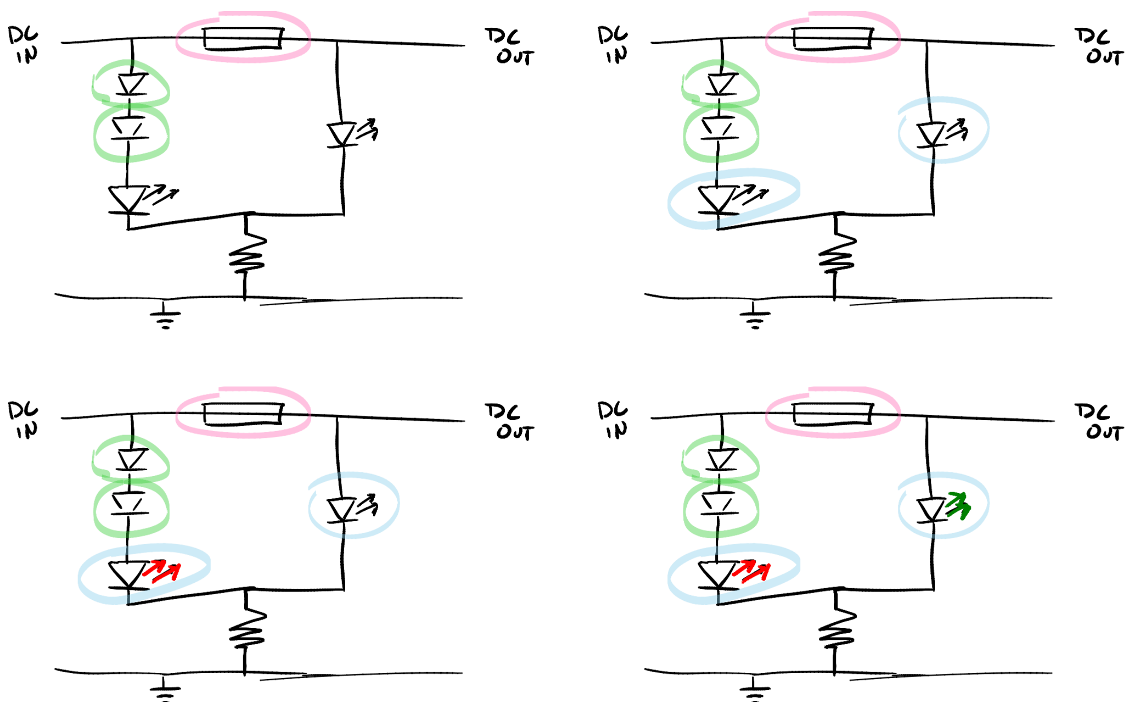


Figure 1-7: User illustrating the different L.E.D.s in Table 1.4.

Our empirical research showed that asking the user a comparison question about the sketch caused the user to add more details graphically and provide a verbal description of the similarities or differences of the components. In this fragment, the system recognizes the two L.E.D. symbols and then queries the user about their similarity. Instead of circling the L.E.D.s the user chooses to modify the sketch and use the ink color to illustrate that the L.E.D.s are different colors.

1.1.5 Example 5

S:	**Speech:** There are two L.E.D.s. Do they have the same function?
U:	**Speech:** Umm, well, no. When the fuse is working, there is less resistance on the right side, ahh... **Sketch:** Draws the current path in green **Speech:** so the green L.E.D. will light. **Sketch:** Draws green light coming from the green L.E.D. **Speech:** If the fuse is blown, then, umm... **Sketch:** Draws blocked path in red **Speech:** the current can only flow through the left side **Sketch:** Draws the current path in red **Speech:** so the red L.E.D. will light up. **Sketch:** Draws red light coming from the red L.E.D.

Table 1.5: User explains which L.E.D. is illuminated.

In our experience short questions about a sketch can trigger long explanations by the user. In this case, asking about the function of two components results in the user describing both operating conditions of the system and the resulting L.E.D. states. Different arrow colors differentiate the two current flow cases. The user also shows the state of the L.E.D.s by using lines to indicate light emission. By asking a simple question, the system can acquire a significant amount of additional verbal and sketch input.

The next chapter describes our initial user study that looked at speech and sketching data. The following chapter describes the details of our more recent study that examined multimodal dialogues and discusses our results and analysis of that study. The subsequent chapter discusses the implications of that study for our system and the challenges we face. We conclude with chapters about related work and our contributions.

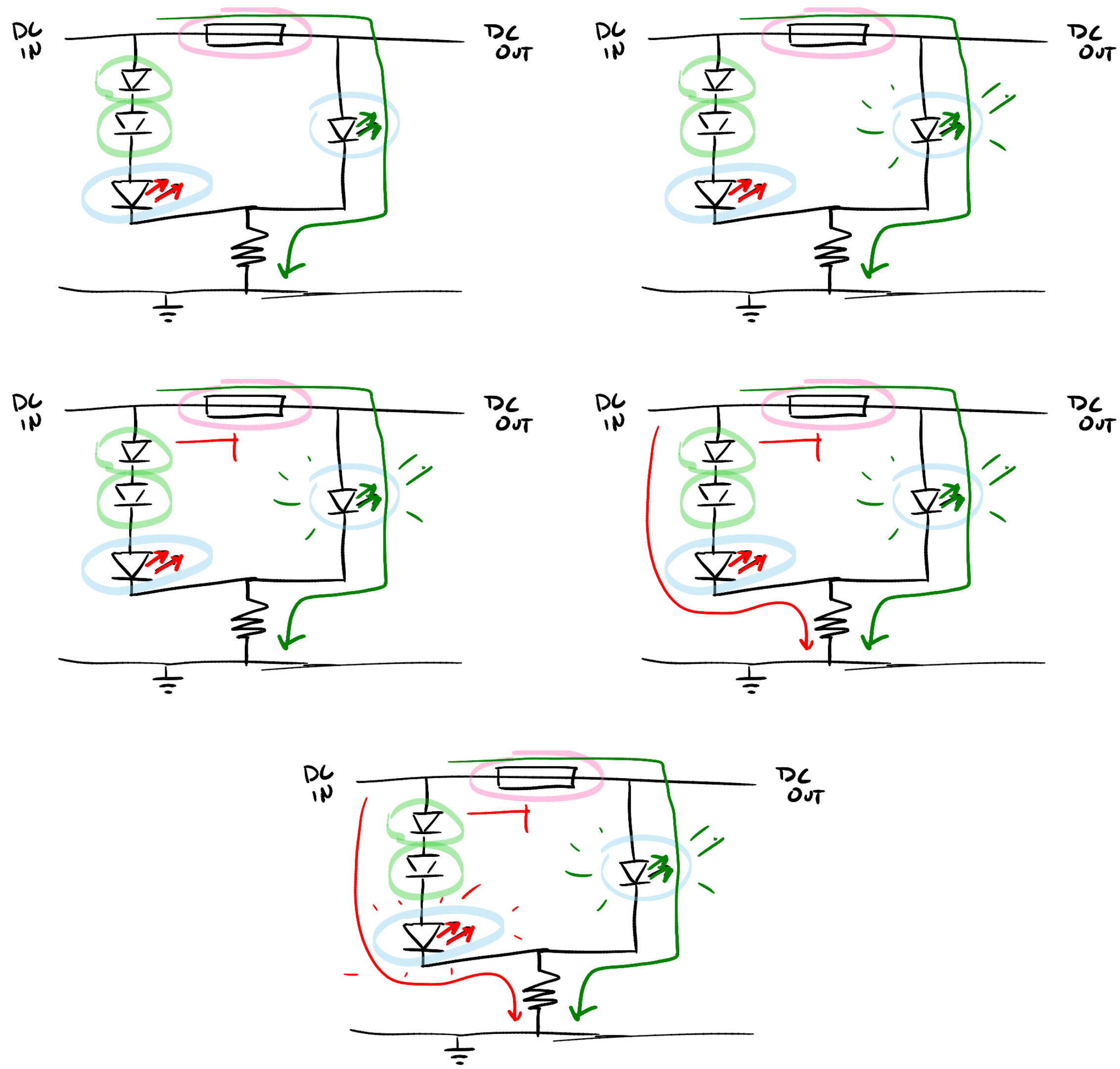


Figure 1-8: User illustrating which L.E.D. is illuminated as discussed in Table 1.5.

Chapter 2

Initial User Study

Our first study was an empirical investigation to collect informal and natural speech from users as they were sketching and verbally describing mechanical systems. The purpose was to identify the vocabulary used to describe mechanical systems and find out which features of the systems were described verbally and which were drawn. In addition, we wanted to identify relationships between the speech and the sketching inputs (e.g., timing, references, ordering) that would enable us to exploit these patterns in the data and create a system that responds to the user's utterances.

Six users drawn from the MIT community were asked to draw and explain six mechanical devices [2]. The users were shown small versions of the devices and then drew enlarged versions of the devices on a whiteboard. The users were videotaped while making their explanations. They were told to describe the devices as if they were talking to a small group of people, such as a physics tutorial. The figures had marks to indicate replicated components and equal distances (Figure 2-1). These graphical marks were provided to get an idea of how the participants would describe identical or equally spaced objects without inadvertently biasing their language by expressing this in words we had chosen.

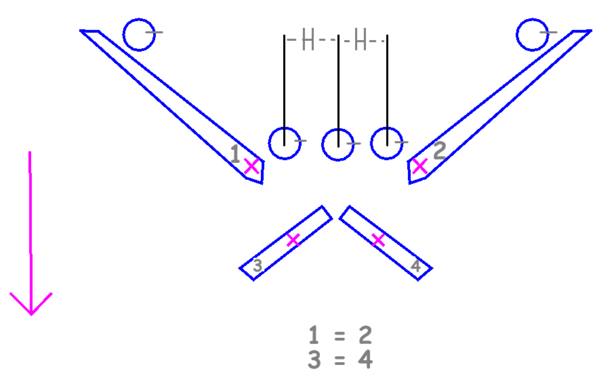


Figure 2-1: One of the devices that the participants sketched. The grey hash marks, grey numbers, and equivalences indicate congruent components.

The data from the videos were analyzed by manually transcribing and assigning timestamps for individual speech events (roughly, phrases) and sketching events (part of a drawn object). Subsequently, topic shifts and corresponding events were manually annotated. For example, three sketched pendulums and the speech describing the three pendulums formed one topic. Speech phrases about different topics were separated. For example, the speech describing objects on a ramp was separated from the speech describing the ramp itself. Observations about the speech and sketching that emerged from the data are discussed below.

2.1 Observations about the Data

Using the manually timestamped, transcribed and grouped data from the user study, we tried to determine what sort of vocabulary and multimodal integration patterns were natural when describing the mechanical systems. Several general patterns emerged from the data.

- Disfluencies, such as "ahh" and "umm" were good indicators that the participant was still talking about the same topic. For example, a participant said: "And then we have this umm [draws rectangle] table." The word "table" occurs after the disfluency "umm" and is the conclusion of the sentence not the beginning of a new sentence.

- Key phrases such as "there are", "and", and "then" were indicators that the participant was starting a new topic. For example, a participant started a new topic beginning with "then" in the following utterance: "Then we have like a [drawing rectangles] divider in that box."

- Consecutive instances of the same drawn shape indicates that the shapes represent the same type of object and that the topic is the same for all of the instances. For example, in a device that contained two pulleys, most participants drew both of the pulleys consecutively.

- A gap (an absence of input) of more than about 0.8 seconds in both the speech and sketching inputs indicated a significant pause by the participant. In the following utterance, the participant separated two topics with a pause: "So now we have a box [draw box] with five circles [draw one circle] inside on the top [draws four circles] [pause] And then we have like a..."

- We observed that participants never talked about one topic while sketching about another topic. For example, participants didn't speak about springs while drawing a ramp.

- Combinations of the above observations reveal other patterns. For example, divisions between topics frequently occur when a pause precedes one of the key phrases. One topic segment might include three sketching events each consisting of a spring and the speech phrase "[pause] And that's ahh filled with springs."

Domain specific vocabulary is also critical for understanding the multimodal input. Linking the noun “pendulum” with the corresponding sketch components – a rod connected to a circular body – is critical to resolving references to “pendulum” or “pendulums.” A deeper understanding of the structure and function of a pendulum is required to act upon references to modifier adjectives such as “identical” or “touching.”

2.2 Overview of Our Initial System

To date our group has developed systems that understand sketches (but not speech) in a variety of domains [4, 12, 24]. For example, ASSIST lets the user sketch in a natural fashion and recognizes mechanical systems, then interfaces with a simulation tool to allow users to view their sketch in action.

Using the observations from the user study, we built a system that could modify a sketch created in ASSIST by combining speech recognition and the sketch interpretations. This enables users to say things like “there are three identical equally spaced pendulums” while sketching several pendulums. The system will then respond by making the pendulums identical and spacing them equally, as shown in Figure 2-2. The system has several components including speech recognition, a rule system, and an integration framework. These components are described in more detail below.

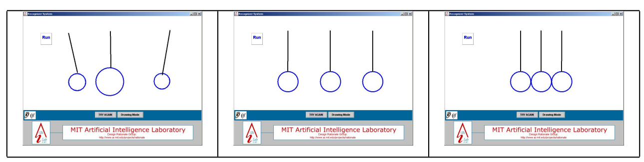


Figure 2-2: Three successive steps in our multimodal system. The first image shows the sketch before the user says anything. The second image shows the sketch after the user says “there are three identical equally spaced pendulums.” The third image shows the sketch after the user says that the pendulums are touching.

2.2.1 Speech Recognition

The vocabulary and sentences from the transcribed videos, augmented with a few additional words (e.g., plurals and numbers), were used to create a speech recognizer for the system. The speech understanding is provided by part of Galaxy[14], a speaker-independent speech understanding system that functions in a continuous recognition mode. The system allows users to talk without prior calibration of the system and without having to warn the system before each utterance. Both factors help create a natural user interface.

2.2.2 Rule System

Awareness of the observations and the patterns described above (Section 2.1) allowed us to build a multimodal topic segmentation system using sketch and speech data.

We manually derived a set of approximately 50 integration rules that encapsulated the knowledge gathered from the user study.

Some rules group objects that are the same shape (e.g., grouping consecutively drawn triangles), others use the timing between the speech and sketching events to identify overlapping events and pauses between events, while others look for key words that we observed were good indicators that the user started a new topic. Other rules identify key times that separate groups of related speech and sketching events. One such rule indicates a possible new group when a speech utterance starts with a key word that is preceded by a pause. The rules were created using 18 data sets. The rules were kept general and do not use specific features or vocabulary of the mechanical engineering domain.

The result of the rules is a determination of the key times which delineate groups of speech and sketching events that refer to the same objects. For example, this might produce a group that included two sketched springs and the speech phrase "that's suspended by springs on the bottom."

2.2.3 Integrating Speech and Sketching

Integrating the speech into the sketching framework allows the user's utterances to affect the sketch. There are three stages to the processing of the speech and sketching. The initial partitioning of both is done by the rule system. In contrast to the rule system, the subsequent phases of the integration make use of domain specific vocabulary. In the second phase, a search is conducted within a group found in the first phase to align the speech and sketching events (e.g., match the speech event containing the word "pendulums" with any sketched pendulums). In the third phase, the search is widened to adjacent groups in the event that the correspondence can't be found in the original group alone. The third phase relaxes the constraints determined by the rules to provide more flexibility in the grouping and accounts for domain specific vocabulary.

The system has a grammar framework that recognizes certain nouns and adjectives and thereby produces a modest level of generality. For instance, one noun it can recognize is "pendulum." The system needs to be told what a pendulum looks like, i.e., a rod connected to a circular body, so that it can link the user's intentions (e.g., drawing three identical pendulums) to a modification of the sketch. Adjectives it can recognize include numbers and words like "identical" and "touching." Adjectives are modifications to be made to the sketch (e.g., "touching"). The framework is general enough to allow the system to be extended to work with more examples.

For example, for Newton's Cradle, functions were needed to space the pendulums equally and to make them identical. Changing the sketch required performing a simple translation from the descriptions, such as "equally spaced," to a set of manipulation commands that were implemented in ASSIST. Figure 2-2 illustrates one possible interaction that results in a modification of the original sketch.

This process of segmenting and aligning the data also allows us, in a limited way, to use both modalities in interpretation. For example, if the user draws three pendulums and says there are two, the system will ignore the speech. However, if the user says that there are four pendulums, then the system will wait for another pendulum to be drawn before attempting to group the speech and sketching events.

2.3 Results

To determine how well the rules work, the transcript files from the videos were parsed and run though the rule system, with each speech and sketching action presented sequentially as if arriving from a user. The data used to test the system was separate from the data used to create the rule system.

The results of running the rules on the video transcripts were compared in detail to hand-generated results for 4 data sets that comprised the test set. There were 29 topic separation times in the hand-generated segmentations. The computer-generated segmentation matched on 24 of these, and found 18 additional separation times. The 18 additional times were analyzed by hand and further classified as “incorrect,” “inconsequential,” or as resulting from “shallow knowledge.” The “inconsequential” category includes separation times that were immaterial to parsing, such as separation times added at the beginning, prior to any speech or sketching events, and extra times between some speech events at the end of the interaction (see Table 2.1). For example, at the end of the interaction a participant spoke the five speech fragments in Table 2.1. The hand segmentation placed all five events into the same group, however, the software placed the events into three groups by placing “inconsequential” topic segmentation times between speech events 1 and 2 and between speech events 3 and 4. The “shallow knowledge” category contains additional topic separation times that were placed between sketching events (see Table 2.2). The hand segmentation placed all five events into one group, but the software placed the events into two groups placing an extra topic segmentation time between events 3 and 4. The rules do not have any knowledge of the meaning of the anchor or the spatial relationship between the ramps. As a result, the rules did not place these events into the same group, as the hand segmentation did.

1	“I’m puzzled as to how to indicate that”
2	“equal size of”
3	“the suspended balls”
4	“and that it is not the same as”
5	“the falling balls”

Table 2.1: Data from one of the participants exhibits how the speech we are working with is not grammatical.

The hand segmentation had the advantage of having all the sketching and speech events to examine at once, as well as the spatial relationships between sketched components. The software segmentation processed speech and sketching events sequentially and did not have access to any spatial relationship information.

1a	"The slopes are fixed in position"
1b	[draws middle ramp]
1c	[draws middle ramp anchor]
2a	[draws bottom ramp]
2b	"slope"

Table 2.2: Example of a "shallow knowledge" topic segmentation time.

2.4 Limitations of the system

The speech and sketching system worked well for simple cases, but it is limited in several ways. For example, a user could sketch three pendulums and say that they are identical; the system would then make the pendulums identical by averaging their dimensions appropriately. The first shortcoming is that the system could improve the sketch only if the speech and the sketch were both interpreted correctly. If, for example, the user drew and talked about three pendulums, but the system only identified two of them in the sketch, it couldn't edit the sketch. Likewise, if the system found three pendulums but misunderstood the speech and heard a different number, it couldn't do anything.

The second shortcoming is revealed if the user refers to three pendulums, but draws four. The system could only communicate in one direction – listening to the user – it could not ask the user any questions. This left the system with no way to cope when information from different input modalities conflicted.

The shapes and allowable manipulations were hand coded, as were the words that were associated with them. Providing this information was time consuming, which is the the third shortcoming of the system. For example, for pendulums to be "identical," the balls need to be the same diameter, the rods need to be the same length, and they need to be connected the same way. The angle of the pendulums is also important as is the starting location of the rods. How the pendulums should be aligned and constructed is specific to the pendulums and would be different for other shapes.

The system also had no knowledge of the spatial relationships between different sketched shapes such as shapes drawn inside other shapes. For example, the system didn't know that a particular anchor was drawn inside of a ramp. Knowledge of these spatial relationships will be necessary for recognizing and modifying more complex sketches.

We want to improve the system both by making communication with the user bidirectional and by providing the system the ability to learn new speech and sketching vocabulary. This will allow the system to enter into a conversation with the user to resolve ambiguities and ask the user questions about their design. The desire to have bidirectional communication was the impetus for the dialogue study described in the next chapter.

Chapter 3

Dialogue User Study and Analysis

3.1 Dialogue User Study

The goal of our work is to build a multimodal interactive whiteboard that can communicate bidirectionally. In order to examine the nature of such an interaction, we conducted a user study to look at natural bidirectional communication between two people. We set up an extensive software architecture to allow two users to sketch simultaneously while sharing the same drawing surface. The intent was to determine things like: what are the characteristics of bidirectional interaction; what questions are asked; how is the sketching surface used to ask questions; how to learn new, out-of-vocabulary terms; how to handle disfluency; how prosody reveals cues about the speakers intentions; how conversations are structured; and how often and when it is okay to interrupt the user.

3.1.1 Motivation

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 using one or two words, not natural speech.
- Unidirectional communication – The system cannot ask questions or add things to the sketch.
- Annotation instead of drawing – The user can only annotate an existing representation, not use free form drawing.
- Fixed set of graphical symbols – The user has to know the fixed symbol vocabulary.

Ideally, we would have conducted a Wizard-of-Oz study in which responses to the participant would appear to be coming from a computer. We determined that this was too difficult, given the open-ended nature of the speech and sketching in the study, and instead used the study protocol described below.

3.1.2 Study Setup

Eighteen subjects participated in the study, all of them students in the Introductory Digital System Laboratory class at MIT. The experimenter and participant sat across a table from each other (Figure 3-1), 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 PC was 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.

This 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 (Figure 3-2). 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.

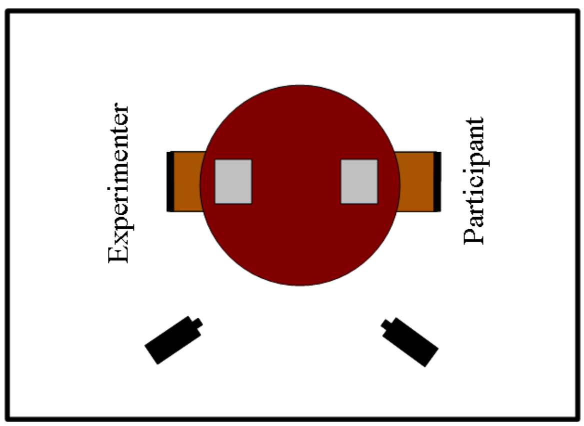


Figure 3-1: Overhead layout of the user study.

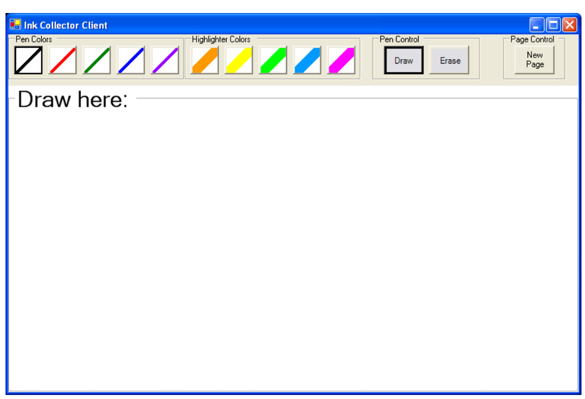


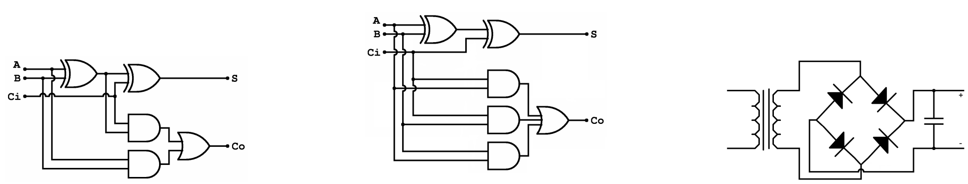
Figure 3-2: The window where the users sketched.

The software recorded the (x, y) position, time, and pressure data for each point in every stroke 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, with the audio and video synchronized. Each camera was connected to a hardware-based video encoding card and the audio/video stream digitized using MPEG2 compression (720x480 pixels, 30fps).

We developed software to provide a variety of services, including ensuring that the timestamps for the sketch data were synchronized with the audio and video data, gathering data about study participants by using a computer-based questionnaire, and displaying instructions. Having synchronized data streams allowed us to replay the study as it happened and facilitated analyzing the timing of the speech and sketching events. To ensure participant anonymity the data was anonymized using a random number.

Participants were instructed to sketch and talk about four different items: a floor plan for a room or apartment 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 text description of the circuit and a list of suggested components. They had the option of viewing a schematic of the transformer or adder circuit (Figure 3-3) before they began drawing, but the schematic was not visible while they were drawing.



(a) Full Adder Schematic I

(b) Full Adder Schematic II

(c) AC/DC Transformer Schematic

Figure 3-3: Schematic views of the full adder and the AC/DC transformer that the participants could choose to view.

At various points in the study, the experimenter added to the sketch and asked questions about different components in it. The participants were compensated with a movie gift certificate valued at $10.

3.1.3 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 for each page of sketching. The XML files contained a full record of the sketching by both the participant and the experimenter. We created software that replayed the study from these data streams.

The software also allowed us to select parts of the audio tracks for playback and transcription. This transcript was passed to the Sphinx speech recognizer [15] 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.

3.2 Study Analysis

3.2.1 Study Statistics

Data from 6 of the 18 participants have been processed as described above. To date data from only six of the participants has been analyzed, due to the time consuming nature of the transcription process. Each of the six datasets contains data from each of the tasks (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. Cumulatively the six participants drew 3206 strokes, 74 erase strokes, and spoke 10,848 instances of 1177 words. Cumulatively the experimenter drew 156 strokes, 3 erase strokes, and uttered 2282 instances of 334 words.

The participants varied in age from 20 to 22, with an average age of 21. There were 14 male participants and 4 female participants. Fifteen of the participants were right handed. Two of the participants owned Tablet PCs, 11 reported having tried one, and 5 reported never having used one.

3.2.2 Initial Results

Our analysis of the study has focused on how speech and sketching work together when people are interacting with each other. Figure 3-4 shows one of the sketches, while Figure 3-5 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 3-6 the sketch contains the AC/DC converter and strokes indicating the flow of current through the circuit in each of two operating conditions. In addition, there are several highlighter strokes used to identify components in the circuit.

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 organize the observations and our discussion.

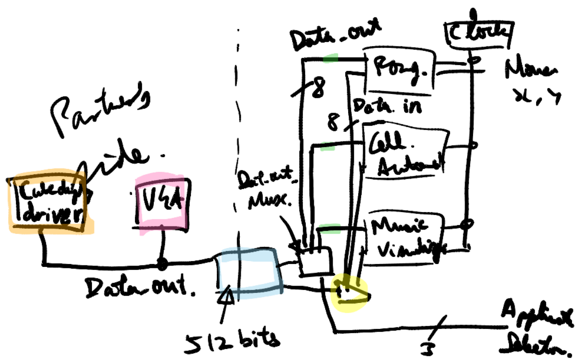


Figure 3-4: A sketch from the user study of a participant's project.

Experimenter:	so all these outputs are are they all the same these outputs
Participant:	um they're not the same they are the individual um um data out connectors of each of the different um well actually i shouldn't be drawing that that at all

---

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 3-5: Three fragments of the conversation about a participant's project (Figure 3-4). Notice the disfluencies and repeated words (discussed in Section 3.2.4).

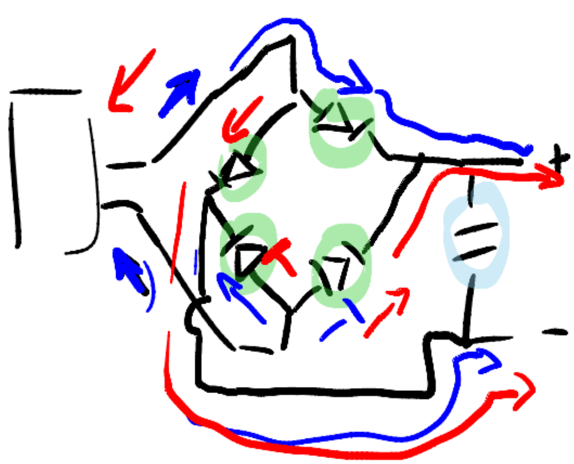


Figure 3-6: A sketch from the user study of an AC/DC transformer.

3.2.3 Observations about Sketching

We found that color was used in several different ways in the sketches:

- To identify regions that were already drawn
- To differentiate objects
- To add an "artistic" character

Identifying Regions

Color was frequently used to refer back to existing parts of the sketch and/or to link different parts of the sketch together. In Figure 3-7, color was used to indicate the location of rooms on a lower floor of the building. In Figure 3-8, three different colors were used to indicate the correspondence between different parts of the sketch – the labeled inputs in the left part of the sketch are highlighted with the same color as the numeric input values in the right half of the sketch. In Figures 3-7 and 3-8 color was critical for identifying references to or connections between parts of the sketch.

Figure 3-7: Yellow highlighter was used to highlight locations of rooms on another floor in a sketch of Next House dormitory.

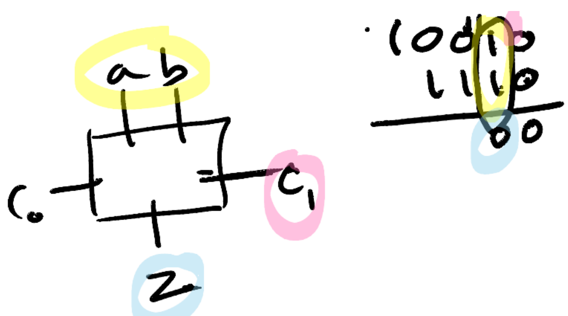


Figure 3-8: Color was used to indicate corresponding areas of the sketch.

Differentiating Objects

Although some participants switched colors while drawing a circuit, different colors were used more often in drawing floor plans, where they were used to differentiate items. When it does happen, the change in color is an excellent indication that the user is starting a new object. This information would greatly aid sketch segmentation. Figures 3-9 and 3-10 are clear examples of a switch in color used to distinguish objects.

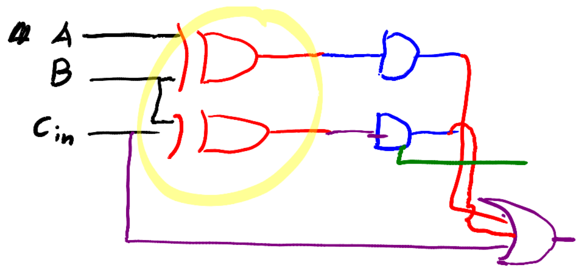


Figure 3-9: Color was used to differentiate the circuit components.

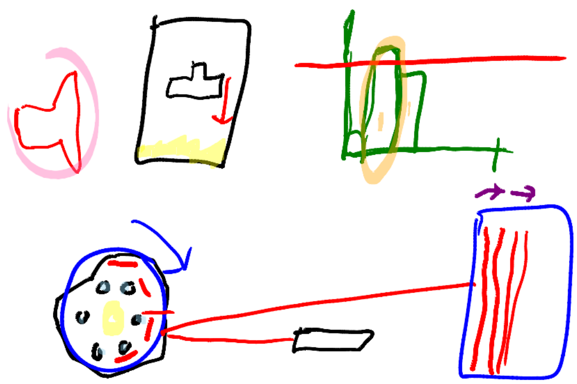


Figure 3-10: Notice that each item in the sketch is a different color.

Adding Artistic Character

The lower-right corner of Figure 3-11, a sketch of a flight simulator, illustrates an artistic use of color. In this case, the user was describing the operation of the attitude indicator. The lower part is brown, indicating the ground, while the upper part is blue, indicating the sky, just as in real attitude indicators. In other sketches, participants used blue to indicate bodies of water, imitating the color of the real world.

Colors used like this still aid in segmenting the input, but also have deeper meaning because they relate to real-world objects and associations. Matching the colors with the references in the speech is one way the system can make connections between the two inputs. For example, one participant drew a blue rectangle in his floor plan sketch and referenced it by referring to the color: "this one's blue [sic] is a sink."

3.2.4 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 fre-

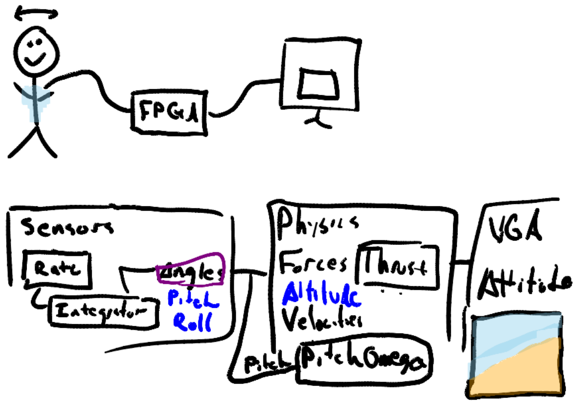


Figure 3-11: Notice the artistic use of blue and orange in the square in the lower-right of this sketch.

quent 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 what is currently being sketched. We address each of these observations in turn.

Disfluent, repetitious speech

The repetition of words or phrases in the speech occurred more frequently when participants were thinking about what they wanted to say. One participant who was 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 better assist the user by asking questions or making suggestions.

Question responses

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 learn to expect a response to questions to have phrasing similar to the question, facilitating the speech recognition task.

Speech relates to current sketching

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 occurring at roughly the same time.

3.2.5 Multimodal

This section discusses three varieties of multimodal interactions between the speech and sketching inputs exhibited by the study subjects: referencing lists of items, referencing written words, and coordination between input modalities.

Referencing lists of items

Participants in the study would often verbally list several objects and sketch the same objects using the same order in both speech and sketching. 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.

Referencing written words

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 [17]. 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.

Coordination between input modalities

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.

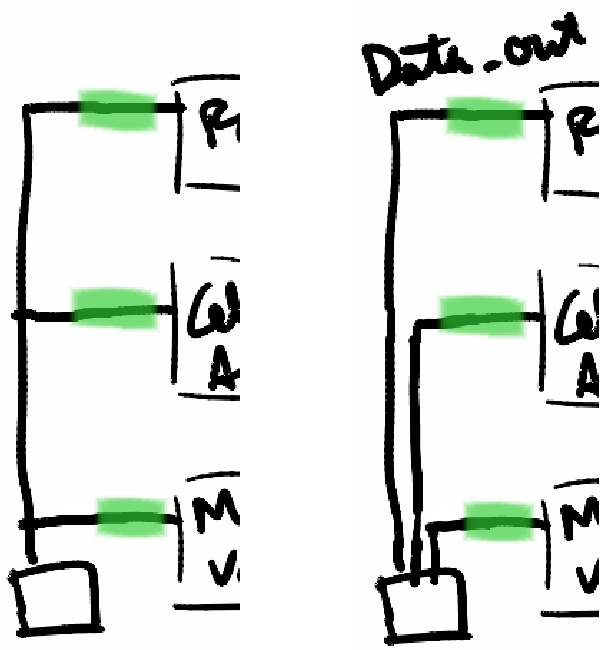
There were many examples in the study where the participant paused their speech to finish drawing an object, then continued talking. For example, one participant said “and that’s also a data out line” and then finished writing “Data out” before continuing the speech. In another case, a participant said “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.

3.2.6 Questions

When the experimenter asked the participants questions, the participants made revisions or explained their design in more depth. This section describes the types of responses participants gave.

Revision

Some questions caused the participant to make the sketch more accurate. Consider Figure 3-12(a), when the experimenter asked if the three outputs, highlighted in green were the same, the participant realized that the original sketch was inaccurate, prompting him to revise it by replacing one data output line with three separate lines (Figure 3-12(b)).



(a) Original (b) Revised

Figure 3-12: The left image shows a piece of the original sketch and the right image shows it after the revision. Notice one data output line on the left side of the original image has been replaced by three lines in the revised image.

Broader explanation

Questions about one part of the sketch also spurred explanations about other, unrelated parts of the sketch, 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 explained the other four labels in the table as well.

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. One participant's reply was that both were AND gates, while another indicated that one was an AND gate and one was 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 to understand 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.

3.2.7 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, but don't relate directly to what they are sketching. Recognizing the uncertainty or other comments could help create a more natural interface for the users.

Another observation from the study is that both 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, while another participant filled in a word that 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.

3.3 Quantitative Analysis

Work in [22] 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 studies used a click-to-talk model for the audio input, but further work showed that this did not affect the results.

We performed a similar analysis of the data from our study. Using the sketching and transcribed speech data for the participants, we matched 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 the results from [22]. The overlap for the *word groups* matches the results in [22], while the results for the *phrase groups* do not.

We compared the start time of the sketching with the start time of the speech, and compared the end time of the sketching with the end time of the speech. Table 3.1 shows the nine possible ways the speech and sketching can overlap and the percentage of time each occurred for the phrase groups. Table 3.2 shows the same thing for the word groups. The enumeration of overlap possibilities is the same as in [22].

Speech Precedes (79%)	Sketch Precedes (19%)	Neither Precedes (2%)
Sketch Speech (0%)	Sketch Speech (1%)	Sketch Speech (0%)
Sketch Speech (49%)	Sketch Speech (6%)	Sketch Speech (0%)
Sketch Speech (29%)	Sketch Speech (13%)	Sketch Speech (2%)

Table 3.1: The temporal overlap patterns for the phrase groups. The alignment of the speech and sketching is shown in an illustration in each table cell. The percentage of phrase groups in each category is also noted.

Unlike that videotape analysis used in [22] to determine the overlap between speech and sketching, we have precise timing data for our speech and have timestamped points from the pen input, both measured in milliseconds. By analyzing the video of several speech/sketching groups whose overlap difference was very small, we determined that 50 milliseconds was a reasonable threshold to use for calling two events simultaneous. We watched the videos of these nearly simultaneous events and could not tell that the speech and sketching started at different times when the difference was 50 milliseconds or less. The video was recorded at 30 frames per second which is approximately one frame every 33 milliseconds.

The graphs in Figure 3-13 and Figure 3-14 illustrate the overlap between the speech and sketching events groups in our data. The x-axis is the time in milliseconds

Speech Precedes (22%)	Sketch Precedes (75%)	Neither Precedes (3%)
Sketch Speech (0%)	Sketch Speech (1%)	Sketch Speech (0%)
Sketch Speech (2%)	Sketch Speech (15%)	Sketch Speech (3%)
Sketch Speech (20%)	Sketch Speech (59%)	Sketch Speech (0%)

Table 3.2: The temporal overlap patterns for the word groups. The alignment of the speech and sketching is shown in an illustration in each table cell. The percentage of word groups in each category is also noted.

that the start of the sketching preceded the start of the speech. A negative number means that the speech preceded the sketching. Similarly, the y-axis represents the number of milliseconds that the end of the sketching preceded the end of the speech. A negative number here means that the sketching ended after the speech. The words in the corners of the graph give a visual depiction of the overlap of the speech and sketching in that quadrant.

Figure 3-13, depicting time differences for the word groups, shows that in most cases (75%), the sketching precedes the word spoken; these data points are in the right half of the graph. The plot has few groups (2%) in the upper-left quadrant, i.e., very few instances of speech that starts first and ends last. Only 20% of the data points are in the lower-left quadrant, i.e., speech that starts and ends first. The graph further illustrates a dense cluster in the upper right. This represents groups where sketching events precede the speech but the speech ends after the sketching. The data is also tightly clustered near the origin; this shows that sketching occurred temporally near the speech that referenced it.

The results for the *word groups* match the results reported by [22]. They reported that 57% of the time writing preceded speech (our data shows 75%). The most frequent overlap category they had was sketching starting first and ending first; this was also our highest category for the word groups (59%).

We also examined the overlap that occurred in the *phrase groups*, as shown in Table 3.1 and Figure 3-14. The phrase plot shows a different relationship from the word plot. Most of the data points are in the left half of the graph (79%) representing phrases where the speech preceded the sketching. Further, many of the data points are in the upper-left quadrant, representing phrases where the speech started before the sketching and ended after it (49%).

This is the opposite of the data reported in [22], 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 where our study

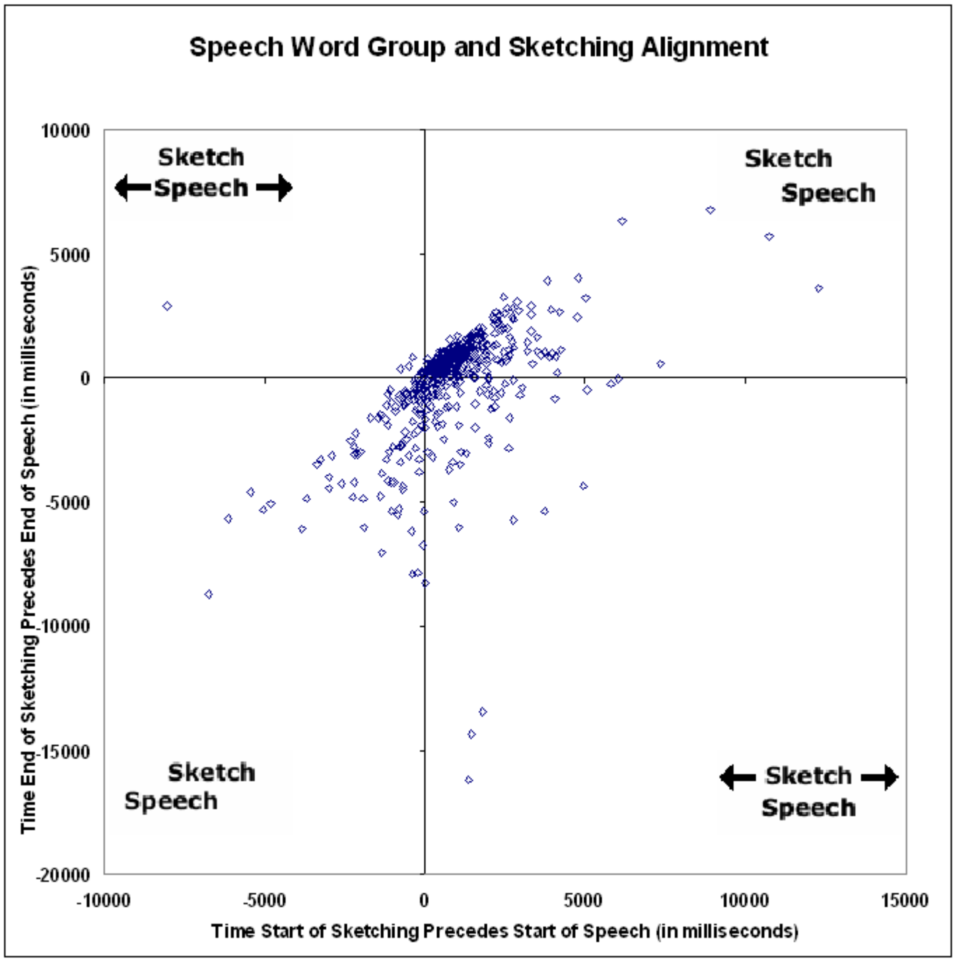


Figure 3-13: A graph depicting the time differences between the start and end times of the speech and sketching in each word group. The x-axis is the time in milliseconds that the start of the sketching preceded the start of the speech. The y-axis represents the number of milliseconds that the end of the sketching preceded the end of the speech. The words in the corners of the graph give a visual depiction of the overlap of the speech and sketching in that quadrant.

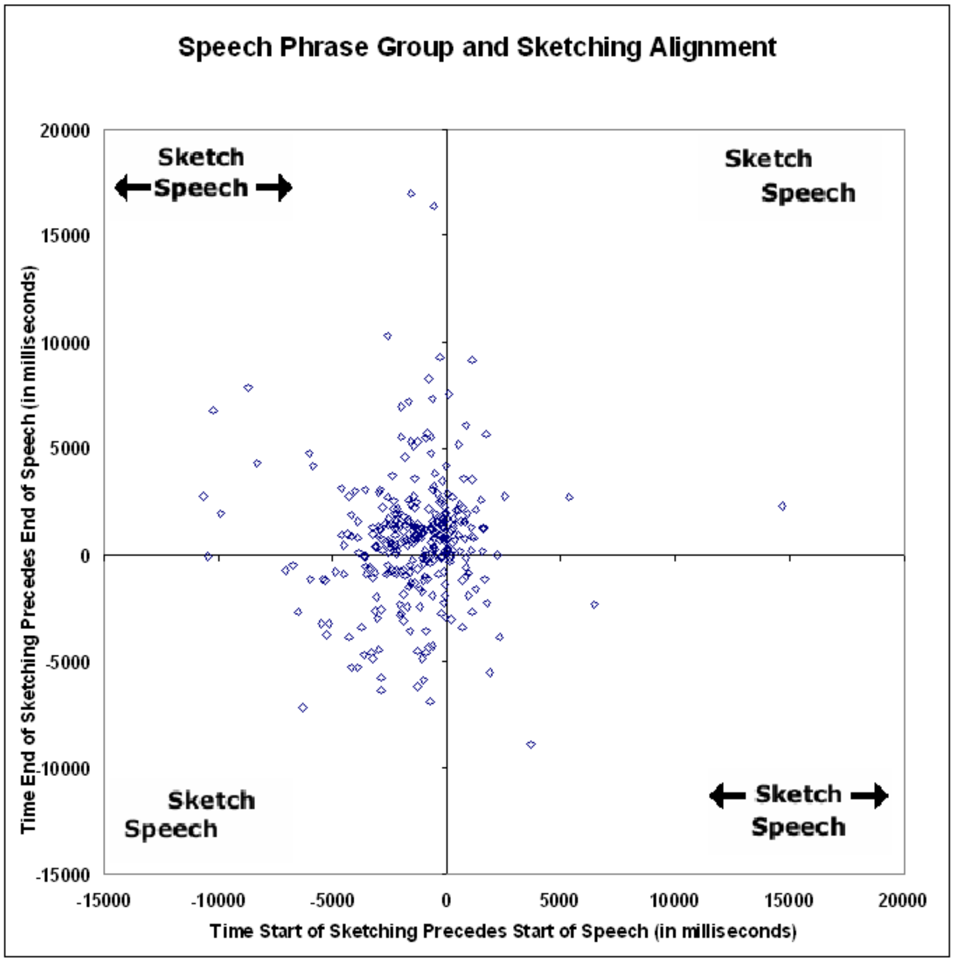


Figure 3-14: A graph depicting the time differences between the start and end times of the speech and sketching in each phrase group.

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. Also, they used a Wizard-of-Oz study, so the participants were talking to a computer instead of a person across the table. The interactive conversation in our study could also have had an effect on the timing of the type of speech and sketching data that was observed.

We tested whether the mean of the difference between speech onset and sketch onset in Figures 3-13 (word group data) and 3-14 (phrase group data) was statistically different from zero. The word data mean different is 795 ms and is significant ($t(495) = 9.93$, $p<.01$); likewise, the phrase data mean difference is -1427 ms and is significant ($t(313) = -10.7$, $p<.01$).

Chapter 4

Resulting System and Evaluation

Our observations and analysis of the dialogue user study provide some implications for the architecture of our multimodal, interactive, digital whiteboard system, and identify some of the challenges we face. We can divide the system into four parts that build on each other:

- Input
- Action
- Dialogue
- Uncertainty.

We discuss each of these parts in turn, along with their architectural implications and challenges. Throughout this chapter we will use several examples to illustrate the types of conversations that occurred during our user study. Each of these examples is pieced together from the user study sketches. In each table, the lines labeled "U" represent user speech and sketching and the lines labeled "S" represent system speech and sketching. Ellipses (...) in the tables indicate pauses in the speech.

4.1 Input

This part of the system combines the speech and sketching inputs together. Ideally, it should not restrict what the user is drawing. The sections below describe the challenges in processing the speech and sketching inputs and how we can approach this problem using timing, disfluences, color, and other properties of the speech and sketch inputs.

4.1.1 Combining Speech and Sketching

The user study revealed patterns in the timing between speech and sketching events, particularly in the timing between naming something and sketching it. Table 4.1 and Figure 4-1 illustrate the overlapping multimodal speech and sketching that was ubiquitous in our user study. 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. Table 4.2 and Figure 4-2 give an example of a user sketching and speaking about a list of objects in a floorplan sketch. A list of objects could help combine the speech and sketching in two ways. The number of spoken items could be used to determine how many sketched objects to expect. Similarly, the number of sketched objects could be used to determine how many spoken objects to expect.

U:	**Speech:** First we take the XOR of A and B. **Sketch:** Draws black XOR gate
U:	**Speech:** That's A XOR B. **Sketch:** Writes A$\bigoplus$B in green
U:	**Speech:** Now we need to take the XOR of that and Ci. **Sketch:** Draws that in blue
S:	**Speech:** Are these the same? **Sketch:** Draws green dot on both XOR gates
U:	**Speech:** Yes. Those are both XOR gates.
U:	**Speech:** So that gives us our sum output. **Sketch:** Draws S and circles it

Table 4.1: Part a sketch of a full adder illustrates a typical example of the overlapping multimodal speech and sketching we observed.

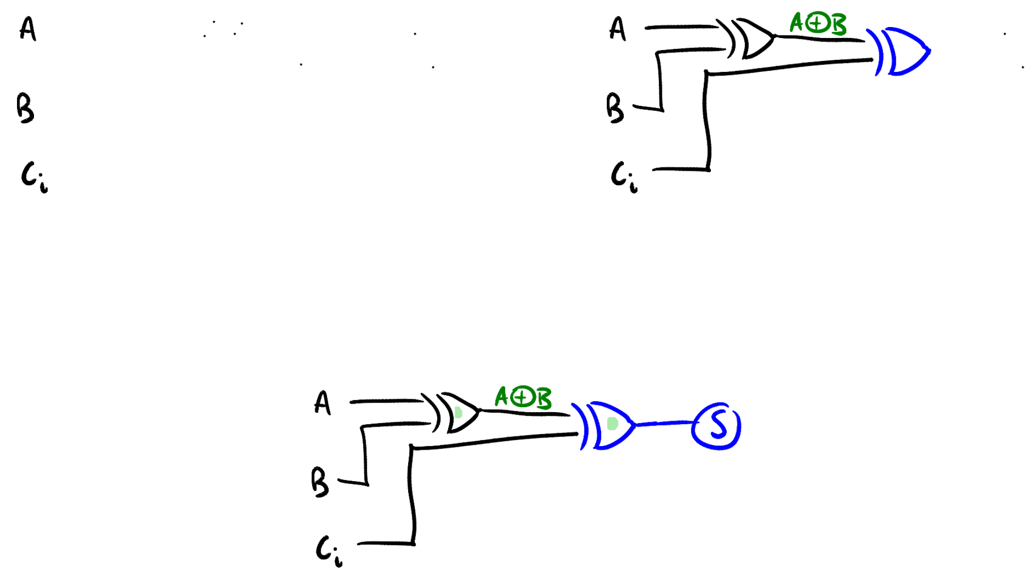


Figure 4-1: The full adder sketch as it is being drawn in Table 4.1. Notice the overlapping speech and sketching.

U:	**Speech:** I've got my desk in the corner...
	Sketch: Draws first desk in green
	Speech: and another desk...
	Sketch: Draws second desk in green
	Speech: and a chair.
	Sketch: Draws chair in red

Table 4.2: Part a sketch of a floorplan illustrates how a user typically spoke and sketched about a list of objects.

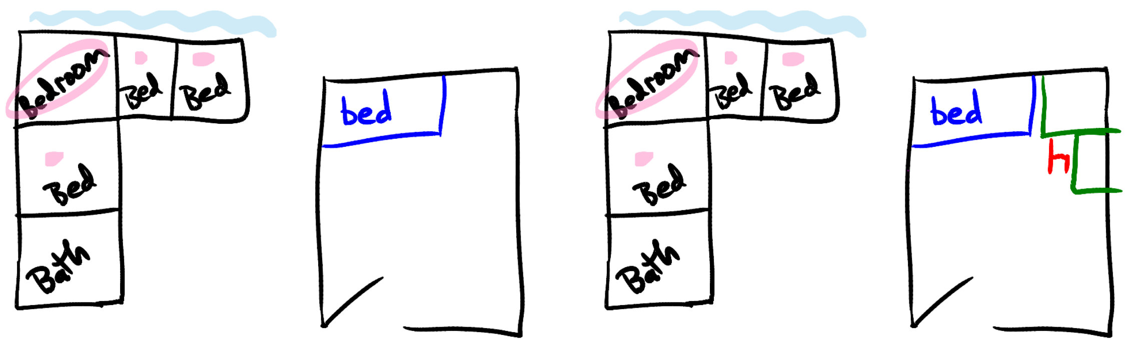


Figure 4-2: The floorplan sketch as it is being drawn in Table 4.2.

The graphs in Figures 3-13 and 3-14 help illustrate that the timing of speech and sketching events is more complex than some previous studies have indicated. Sketching tends to start before speech on a word level, but speech tends to start before sketching on a phrase level. This relationship precludes some of the simple strategies others have used to align speech and sketching, for example, using a time threshold to determine the correspondence between the speech and the sketching.

There are several challenges in combining the speech and sketching inputs. We need a model or representation for the user's speech and we need to find the correlations between the speech and sketch inputs. Previously we used key word spotting with a loose grammar framework. Something similar might be appropriate given the informal nature of the speech we are working with. However, with only key word spotting, the system would run into problems with negatives – for example the user could say "this is not a diode." In addition the system will need some knowledge about the domain. Our goal will not be to handle all possible speech, but rather to handle a subset of it. A more advanced version of the word spotting method we used in our previous system might provide a sufficient level of understanding without trying to solve the difficult problem of natural language understanding.

4.1.2 Disfluencies

Our research so far has shown that disfluencies play an important role. For example, disfluencies in the speech input seem to indicate that the user is still thinking about the same topic. The system can gain an important insight – that the user is still speaking about the same topic – by recognizing the user's disfluencies.

The disfluent nature of the speech will be challenging to recognize. As discussed in Section 4.1.1 and implemented in our previous system, we can accomplish a significant amount without trying to recognize every word correctly. Recognizing repeated words also appears to be advantageous. The speech we observed in the user study at times sounded like the speech in Table 4.3. Figure 4-3 illustrates the sketch before and after the interaction in Table 4.3.

U:	**Speech:** ... So now... umm... now I'm going to... umm I'm not sure how... to compute the carry out. umm... I'm going to draw a truth table... a truth table.
S:	**Speech:** Ok.

Table 4.3: This example from a sketch describing a full adder shows speech that is disfluent and repetitious.

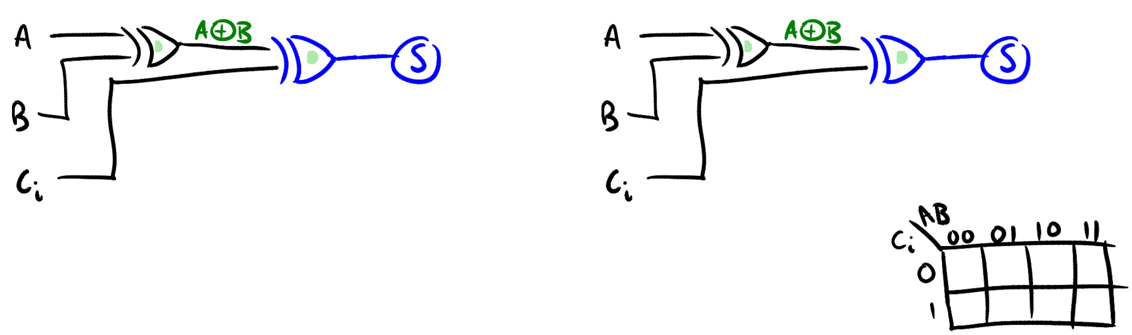


Figure 4-3: The full adder sketch before and after the speech in Table 4.3.

4.1.3 Speech Characteristics

The speed and prosody of the user's speech could be another source of information for the system. Qualitative observations from the user study showed that if a user is speaking quickly they are most likely have a great deal that they want to say about an idea. If they are speaking slowly then they are more likely to be thinking about the design or what to draw next. Topic changes could be detected by observing changes in the speed of the user's speech. It is also possible that different speech speeds indicate different costs for interrupting a user. Interrupting a user when they are speaking quickly may be significantly less desirable than interrupting a user who is speaking slowly.

4.1.4 Color

The dialogue user study revealed the importance of ink color (Section 3.2.3). The system should integrate knowledge about color and recognize that switching ink colors is likely to indicate a new object or new type of object. An example of the user switching ink colors while sketching a floorplan is shown in Table 4.4 and Figure 4-4. Color can also be used as shown in Table 4.5 and Figure 4-5 to show the actual colors of an object. One question that the user study raises is whether the same patterns of color use continue over longer periods of time.

U:	**Speech:** The river is over here. **Sketch:** Draws blue highlighter illustrating the river **Speech:** I can see it from my bedroom.
U:	**Sketch:** Draws outline of bedroom in black **Speech:** This is my room. The door is here. **Sketch:** Draws door in black
U:	**Speech:** This is my bed. **Sketch:** Draws bed and labels it in blue

Table 4.4: In a floorplan sketch the user draws different objects with different ink colors.

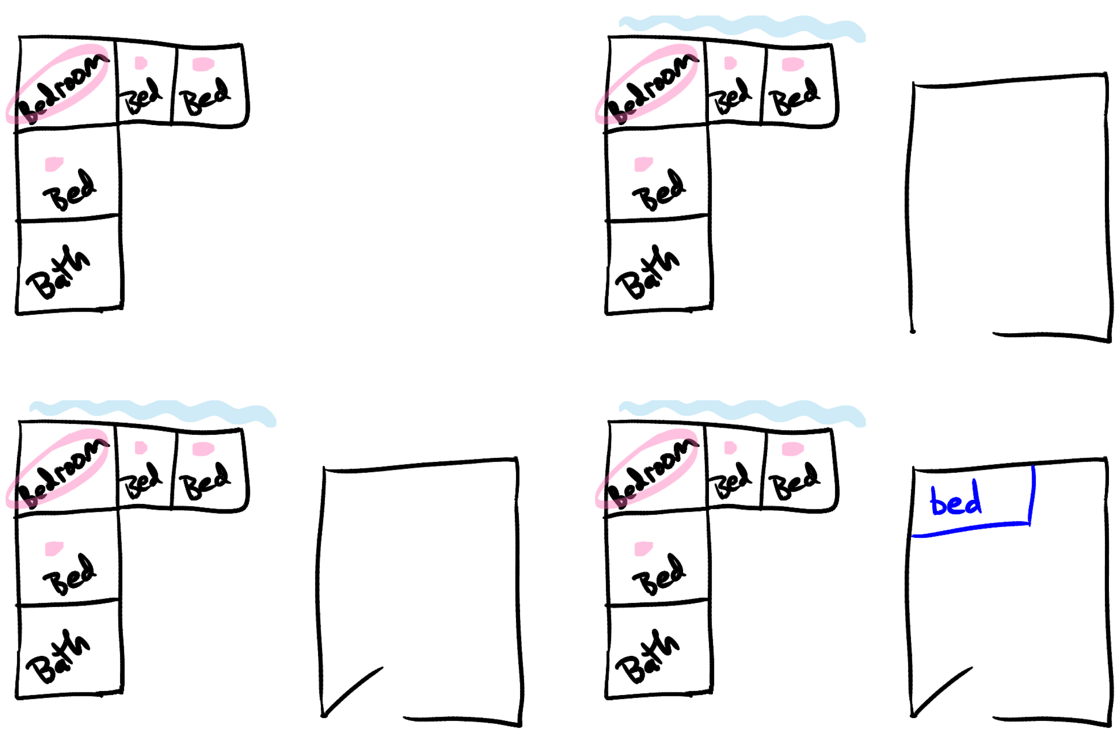


Figure 4-4: The floorplan sketch as it is being drawn as described in Table 4.4 illustrating the use of different colors for different objects.

U:	**Speech:** I also have a projector.
	Sketch: Draws box
	Speech: It has three lens, a red...
	Sketch: Draws red
	Speech: a blue...
	Sketch: Draws blue
	Speech: and a green.
	Sketch: Draws green
	Speech: It projects onto the wall here.
	Sketch: Draws red line

Table 4.5: In a floorplan sketch a projector is drawn with different colors of ink for each of the different lens colors.

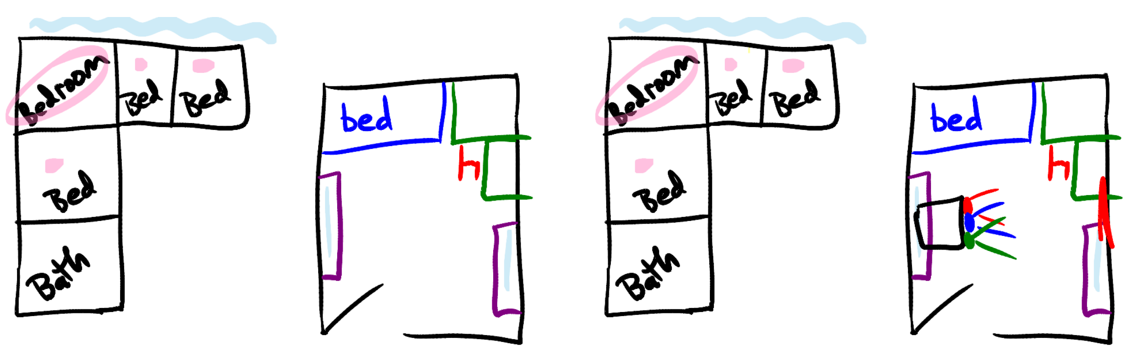


Figure 4-5: The floorplan sketch as it is being drawn as described in Table 4.5. The projector lenses reflect their actual colors.

4.1.5 Handwriting

Users often write words while they are sketching. Components of the system will need to identify which parts of the sketch are handwriting and need to associate handwritten words with their referents. Abbreviations are common and need to be accounted for. For example, “Lookup Table” might be abbreviated as “Look T.” One possible approach to this challenge would be to use pen speed, time between strokes, and spatial information to help separate handwriting from the other parts of the sketch. Handwriting contained within boxes might be less difficult to identify. Handwritten text also occurs next to objects which might be another indicator that the strokes are handwritten words. Table 4.6 and Figure 4-6 are an example how handwriting can be used in a floorplan sketch to identify the types of various rooms.

U:	**Sketch:** Draws outline of five boxes **Speech:** My suite has some bedrooms and... **Sketch:** Labels “bedroom”
U:	**Speech:** a bathroom. **Sketch:** Labels “bath”
S:	**Speech:** What are these? **Sketch:** Draws 3 pink highlighter dots
U:	**Speech:** Those are other bedrooms. **Sketch:** Labels each with “Bed”
S:	**Speech:** Ok. Are those the same as this? **Sketch:** Circles “Bedroom” in pink highlighter
U:	**Speech:** Yes. But they are different sizes.

Table 4.6: This segment of dialogue from a floorplan sketch shows how handwriting can be used to identify parts of sketch. In this case, handwriting identifies the rectangles in the sketch as particular rooms.

4.1.6 Disambiguation

We can use the speech input to help decipher the sketch input and vice versa, also known as mutual disambiguation. We can use hypotheses about sketched objects or text to provide information about what words to look for, and we can use the speech to know what sketched objects to look for. Speech is a rich input modality and information, such as numerical references, can be extracted from it to aid in the disambiguation of the input. For example, the user might indicate the number of objects he is sketching in the speech. The user may also indicate which objects are connected to each other giving us additional information about the sketch. Table 4.7 and Figure 4-7 illustrate an example where the multimodal speech and sketching input might be used to disambiguate each other.

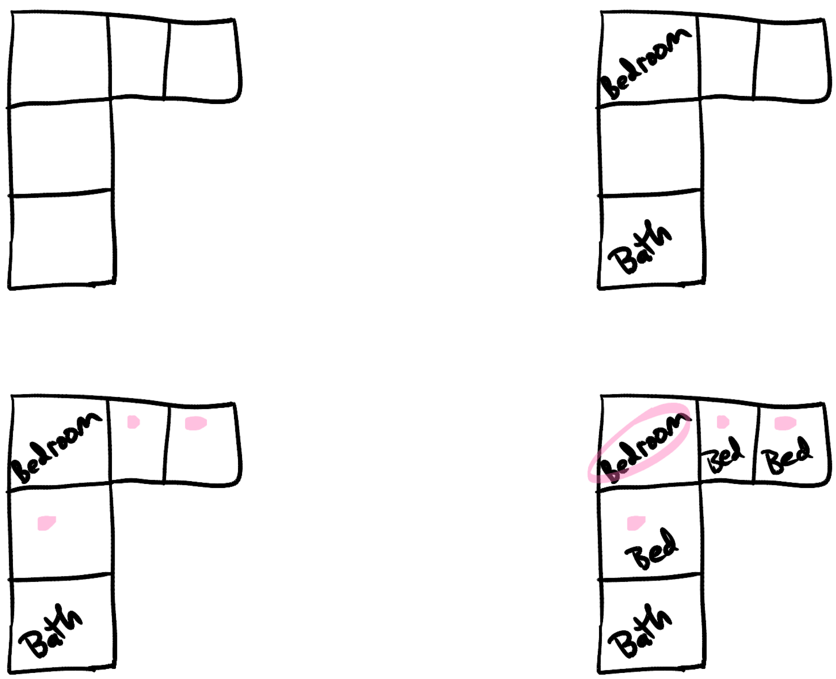


Figure 4-6: The floorplan sketch illustrating handwriting as described in Table 4.6.

U:	**Speech:** This is a bookshelf. **Sketch:** Draws purple box
U:	**Speech:** This is another one. **Sketch:** Draws second purple box
S:	**Speech:** Are they the same? **Sketch:** Highlights both using blue lines
U:	**Speech:** No, they are different sizes.

Table 4.7: This segment of dialogue from a floorplan sketch illustrates how multimodal speech and sketching presents an opportunity to use the speech and sketching inputs to disambiguate each other.

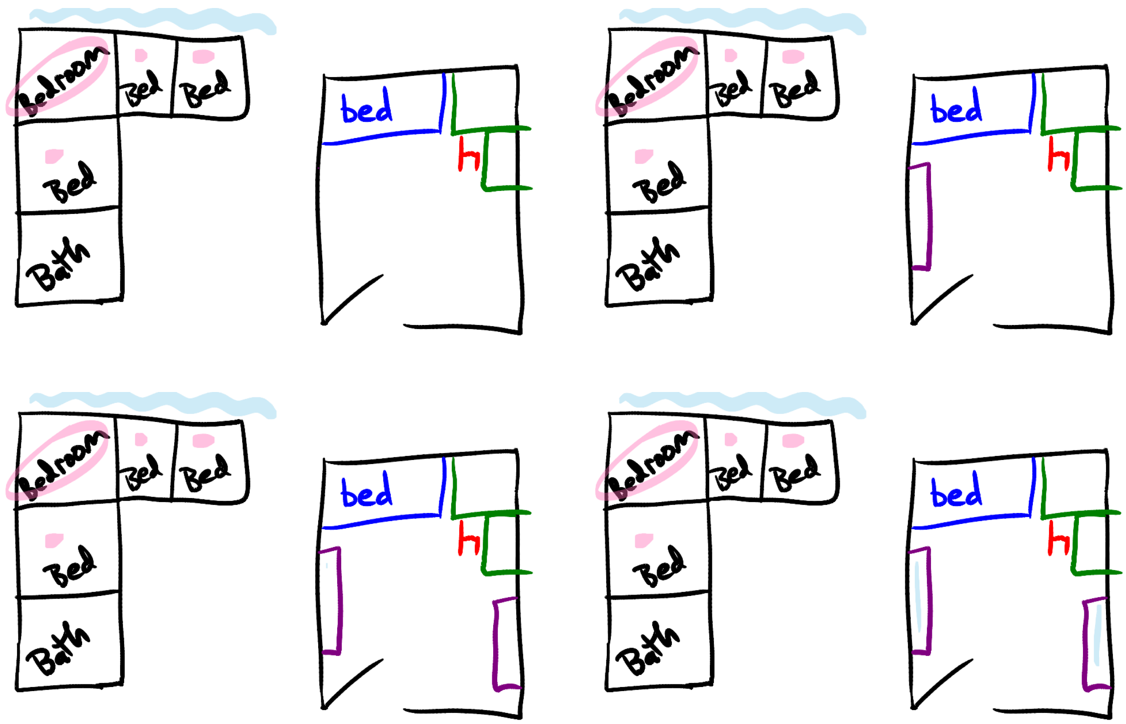


Figure 4-7: The floorplan sketch as it is being drawn as described in Table 4.7.

4.2 Action

This part of the system involves taking some action based on the combination of speech and sketch input. This action might take the form of modifying the sketch based on the user's input or it might be adding new vocabulary to the system based on the inputs. For example, the system might learn that in a floorplan a window is represented as a blue line or it might make a series of windows identical by making them the same length and thickness.

4.2.1 Modifying the Sketch

One problem we face is manipulating the shapes the user draws in a general way. One way to solve this problem would be to base the manipulations on a geometric description of the shapes, such as LADDER [13]. This would avoid the problem we mentioned in Section 2.4 of having to specify how to manipulate each shape individually. In addition, the geometric description would allow us to learn new shapes and vocabulary as described in the next section.

4.2.2 Learning

The system should be able to learn new vocabulary – specifically the correspondence between spoken words and sketched objects. Handwritten words also play a significant role in the sketches we observed in the user study, however, others[18] are making significant progress learning new vocabulary from the combination of handwritten words or abbreviations and the spoken word.

Learning the correspondences between sketched objects and speech would allow the system to extend its knowledge of a particular domain or be extended to a new domain. In the sketches we collected, users would sometimes sketch the same object in different or simplified ways. The system should accommodate this and associate all forms of an object with a word and with each other.

The system can take advantage of the dialogue capabilities discussed in the next section to have a conversation with the user to learn the new word and the corresponding sketch. Interacting with the user will provide an additional means of learning vocabulary that is not available to systems that do not run in real-time (such as Kaiser[18]).

4.3 Dialogue

Our goal is not to have perfect speech or sketch recognition – rather we want to do the best we can with both inputs and get a good idea of the user's intention. By making the system capable of a two way dialogue, we can ask the user for clarifications. Benefits of having a two-way multimodal dialogue include encouraging a rich dialogue with the user and making the interaction closer to the interaction a user would have with another person. Asking the user questions will help keep the user engaged, help the user refine and clarify the design, help the system learn more about the sketch, and make the system more of a partner in the design process. As shown in Table 4.8 and Figure 4-8, the dialogue can cause the user to make corrections to the sketch.

S:	**Speech:** Are all these outputs the same? **Sketch:** Highlights in green
U:	**Speech:** No, actually. **Sketch:** Erases a line and draws it again in purple
U:	**Speech:** One is an eight-bit line and one is a two-bit line. **Sketch:** Adds a purple eight and a black 2 to the diagram

Table 4.8: This example shows the benefit of keeping the user engaged in a conversation – he might realize a mistake in the sketch and correct it.

Building on the basic associations between the speech and sketching inputs, the system will need to generate questions, decide when to ask them, and understand the answers. The questions are also multimodal in that they can involve sketch and speech output. The following sections describe the parts of the dialogue in more detail and outline some of the challenges.

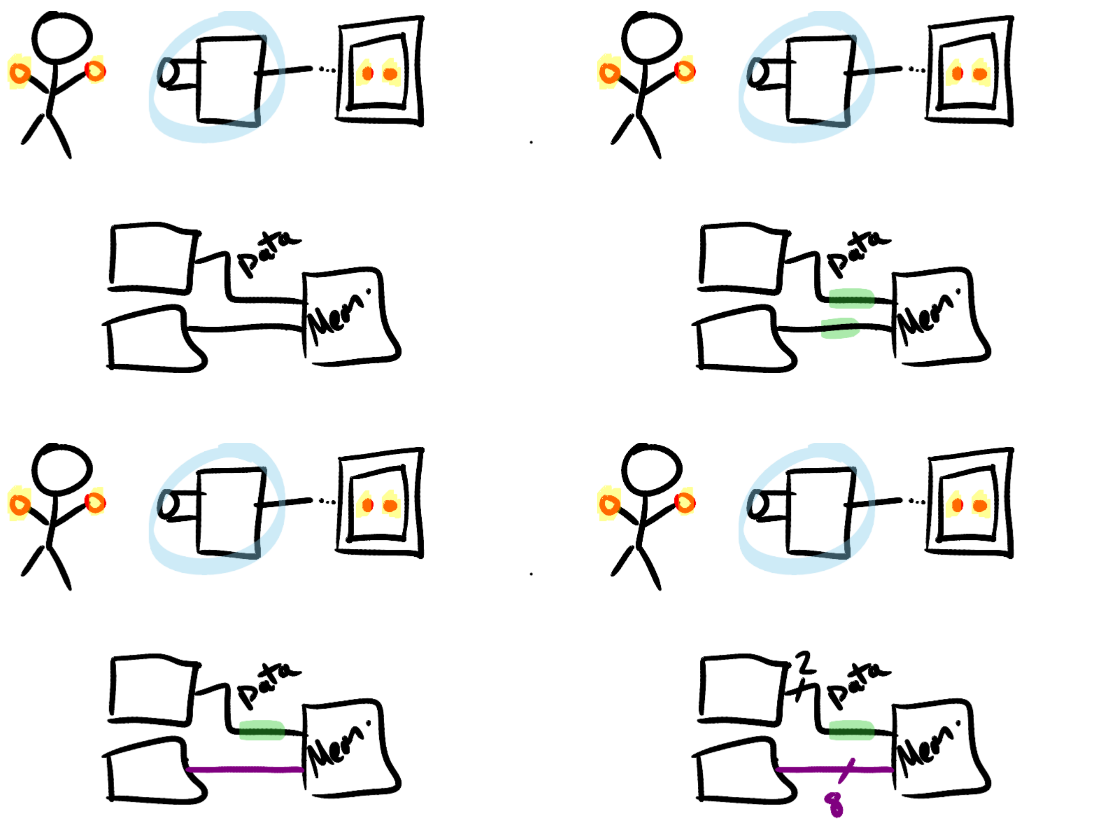


Figure 4-8: After being asked a question, the user revises his project sketch as described in Table 4.8.

4.3.1 Dialogue Management

Spoken dialogue management is an important component of natural user interaction. We hope to build upon the previous research [23, 3, 11], some of which is relevant to managing a multimodal dialogue about design. Ideas such as turn taking and determining which information you still need from the user, are still applicable. The situation is made more complicated by the additional modalities and our desire to make the interaction as natural as possible. In particular, the open ended interaction means that the information the system requires is not known a priori as it is in other domains such as airline reservations. The computer also needs to acknowledge speech by the user, for instance, by saying "ok" as shown in Table 4.9 and Figure 4-9.

U:	**Speech:** So we start with a wave form like this **Sketch:** Draws left wave form in purple **Speech:** and ...
U:	**Speech:** then after the diodes we end up with the absolute value and it looks like this. **Sketch:** Draws middle wave form
S:	**Speech:** Ok.
U:	**Speech:** But it is still bumpy. If you want to make it smoother we add a capacitor... **Sketch:** Draws capacitor
S:	**Speech:** Ok.
U:	**Speech:** and we end up with this. **Sketch:** Draws right wave form

Table 4.9: The system encourages the user to proceed with the explanation with the acknowledgment "Ok."

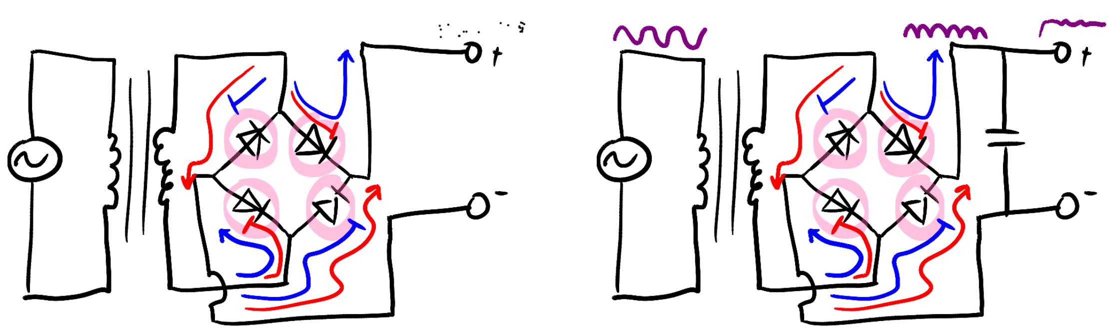


Figure 4-9: The AC/DC converter sketch before and after the description in Table 4.9.

4.3.2 Use of Color in Dialogue

Just as color was useful to the participants in the dialogue user study, using color can be advantageous in the dialogue part of the system. If the system is trying to identify a particular type of object, it could circle the hypothesized instances with a color. If the user was identifying other instances of the objects then they might use the same color. The system could also use different colors for different objects and then the system and the user could refer to the objects using their color. Since color is important we need to make sure the data representation takes it into account. Table 4.10 and Figure 4-10 show how color can be used to show relationships between different parts of a sketch. Table 4.11 and Figure 4-11 show how color can be used to clearly depict different operating conditions of a system.

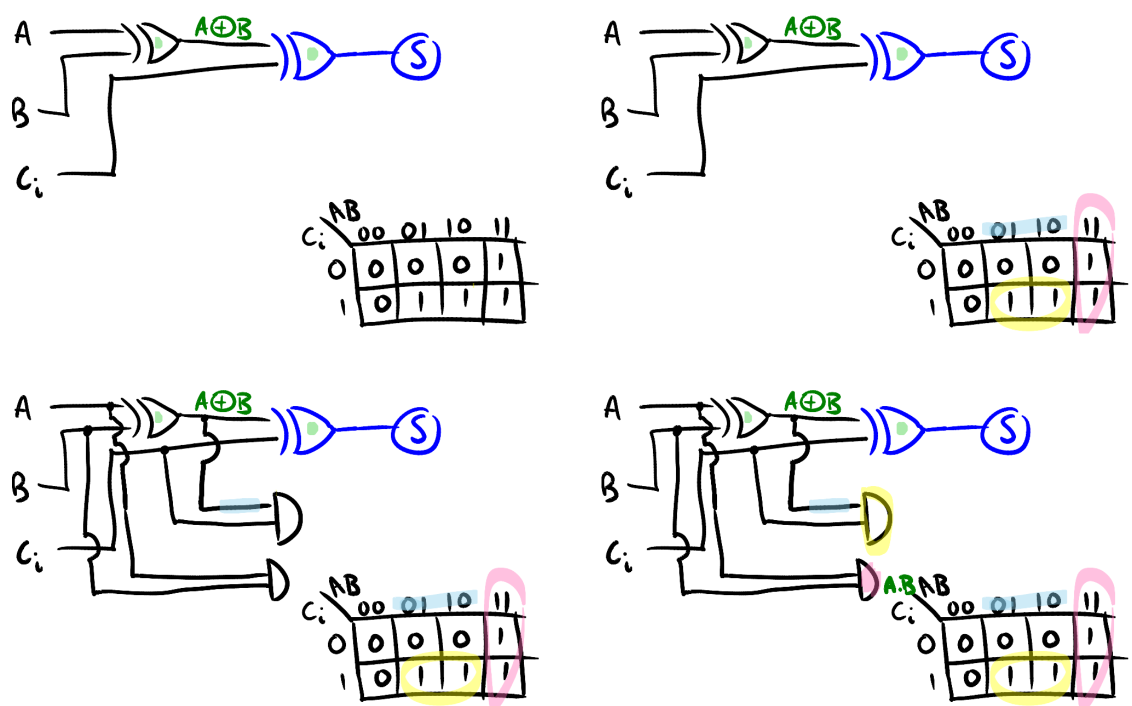


Figure 4-10: The AC/DC converter sketch as it is being drawn as described in Table 4.10. Notice the corresponding parts of the sketch that the colors highlight.

4.3.3 Interrupting

When the system has a question to pose to the user, it needs to decide how and when to ask the question. There are several challenges here: when do you interrupt the user and how confident of the question do you need to be to interrupt the user? In our study, we found that the users could be difficult to interrupt. The system may need a way to alert the user that it wishes to take a turn. Possible cues include starting to sketch, flashing the screen, or saying "umm." An example of the system interrupting a user is shown in Table 4.12 and Figure 4-12.

U:	**Speech:** So now I can see the pattern. We can use one gate here... **Sketch:** Circles in yellow **Speech:** and one here. **Sketch:** Circles in pink
S:	**Speech:** Ok. Where are those gates?
U:	**Speech:** I was just going to draw them. So we have ... these two columns **Sketch:** Highlights in blue **Speech:** which is an XOR. An XOR of A and B.
S:	**Speech:** Ok.
U:	**Speech:** So for the first gate we need A XOR B which we can get from here **Sketch:** Draws wire from A XOR B
S:	**Speech:** Ok.
U:	**Speech:** So this wire **Sketch:** Highlights in blue **Speech:** is A XOR B.
U:	**Speech:** We combine that with Ci **Sketch:** Draws Ci **Speech:** using an AND gate. **Sketch:** Draws AND gate
S:	**Speech:** Ok.
U:	**Speech:** Then we need A and B. **Sketch:** Draws lines for A and B **Sketch:** Draws AND gate
U:	**Speech:** We use another AND gate. So this gate **Sketch:** Outlines in yellow **Speech:** is what I circled in yellow and this gate ... **Sketch:** Scribbles in pink **Speech:** is what I circled in pink. This output is A and B. **Sketch:** Writes A.B
S:	**Speech:** Ok.

Table 4.10: This conversation about a full adder illustrates how color can help show correspondences between different parts of the sketch.

U:	**Speech:** There are two phases to the system. I'll draw the first in blue. The positive current flows this way... **Sketch:** Draws top blue arrow **Speech:** and is stopped from going this way by the diode. **Sketch:** Draws top blue stop **Speech:** On the bottom the negative current flows this way... **Sketch:** Bottom blue arrow **Speech:** and is stopped from going this way. **Sketch:** Draws bottom blue stop
S:	**Speech:** Ok.
U:	**Speech:** Now I'll draw the other phase in red. So the positive current goes this way. **Sketch:** Bottom red arrow **Speech:** And the negative current flows this way. **Sketch:** Draws top red arrow **Speech:** The other two directions are stopped. **Sketch:** Draws two red stops
S:	**Speech:** Ok.

Table 4.11: A user utilizes two different colors to show two current paths in an AC/DC converter.

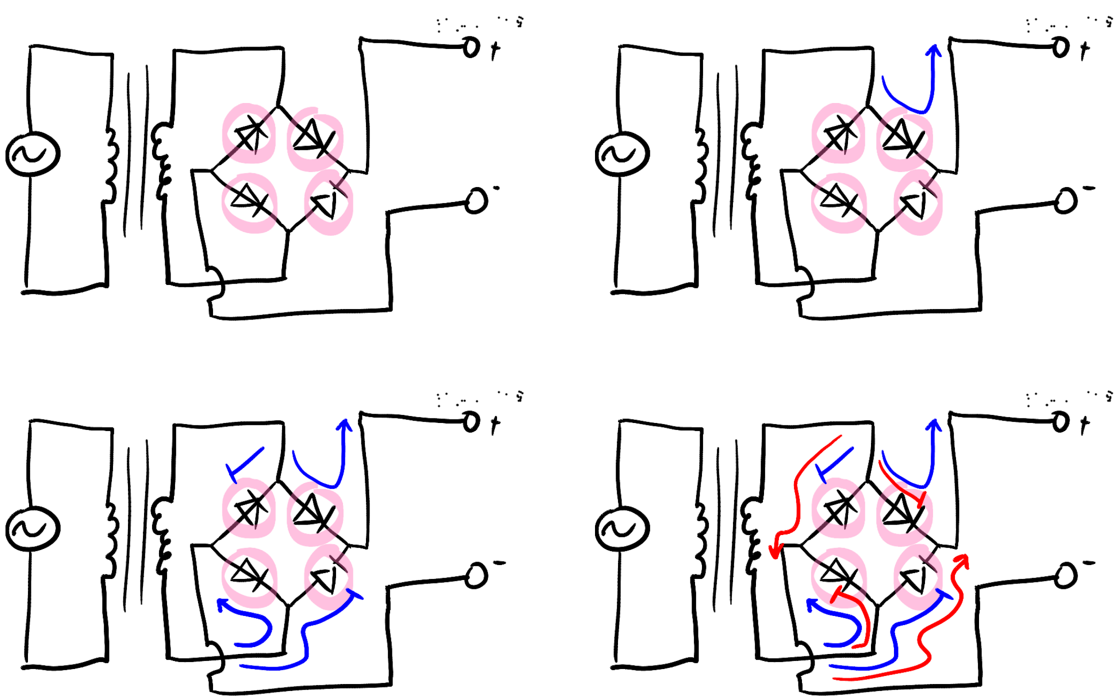


Figure 4-11: The AC/DC converter sketch as it is being drawn as described in Table 4.11. Notice the different current paths illustrated in blue and red.

U:	**Speech:** There are four diodes. **Sketch:** Draws the four diodes
S:	**Speech:** I can only find three diodes. **Sketch:** Circles in pink **Speech:** Where is the fourth one?
U:	**Speech:** That's here. **Sketch:** Circles in pink
S:	**Speech:** Ok.

Table 4.12: The system interrupts the user's description of an AC/DC converter to locate a missing diode.

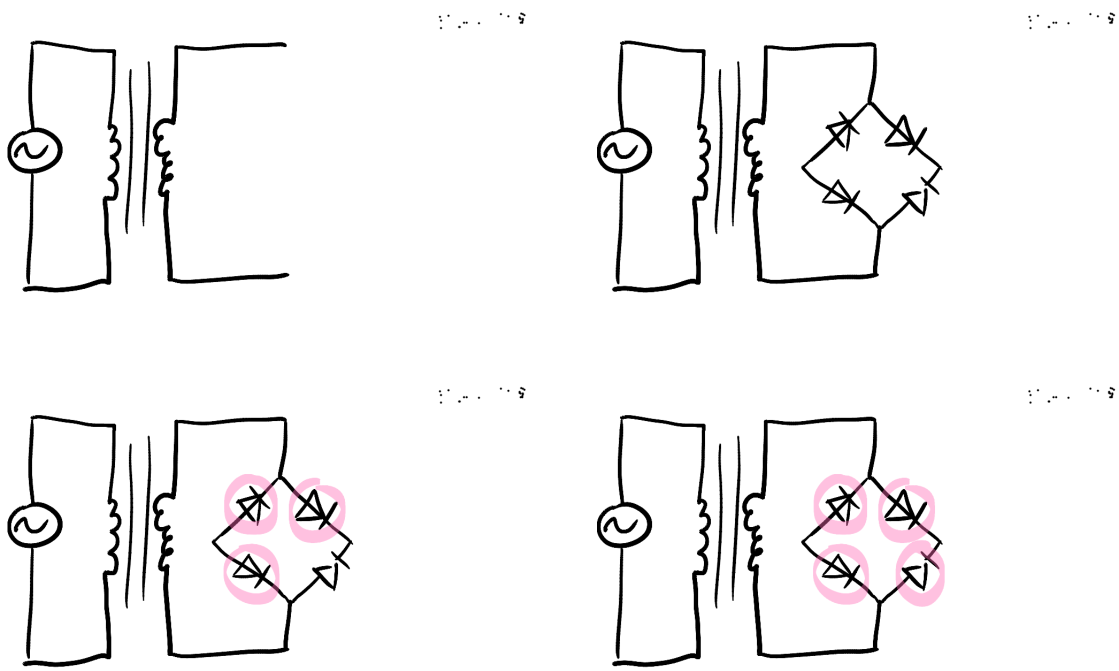


Figure 4-12: The system is uncertain where the fourth diode is located in the AC/DC converter so it asks the user as described in Table 4.12.

4.3.4 Questions

As part of the dialogue, the system needs to be able to ask questions like "Which part is the FPGA?," "Is this the same as this?," and "Where's the diode?" Questions can help the computer resolve uncertainty about the sketch. For example, it might ask "What's this?" or "Is this a diode?" to confirm the identity of a component. Table 4.13 and Figure 4-13 show the system asking the user several questions about their sketch.

There were several times in the user study where a participant couldn't think of the right word and asked for help; the system should be able to attempt to help in these cases too by answering simple user questions.

U:	**Speech:** So now I just need to combine those together. **Sketch:** Connects them with an OR gate
S:	**Speech:** What's this? **Sketch:** Circles OR gate
U:	**Speech:** That's an OR gate.
S:	**Speech:** What's the output? **Sketch:** Highlights output
U:	**Speech:** Oh. That's the carry out. **Sketch:** Writes Co
S:	**Speech:** Ok.

Table 4.13: This piece of dialogue from a sketch of a full adder shows the system asking the user several questions.

4.4 Uncertainty

This phase of the system includes dealing with uncertainty expressed by the user and other comments that the user may make that don't directly relate to the sketch (as described in Section 3.2.7). For example, the user says "now I'm going to switch ink colors." This information can help the system understand the sketch by providing another alignment point between the speech and the sketching, but isn't directly about the sketch. The users also expressed uncertainty about the designs by saying things like "I'm not sure this is right..." A challenge for the system is to take this uncertainty or conflicting information into account.

4.5 Evaluation

Once the system is built, we plan to conduct a user study to measure its performance via several comparisons. First, we can compare an existing sketching system (like ASSIST) to ours to see which one users prefer. Second, we can compare different versions of our system to each other to see how effective the various capabilities are. We can compare:

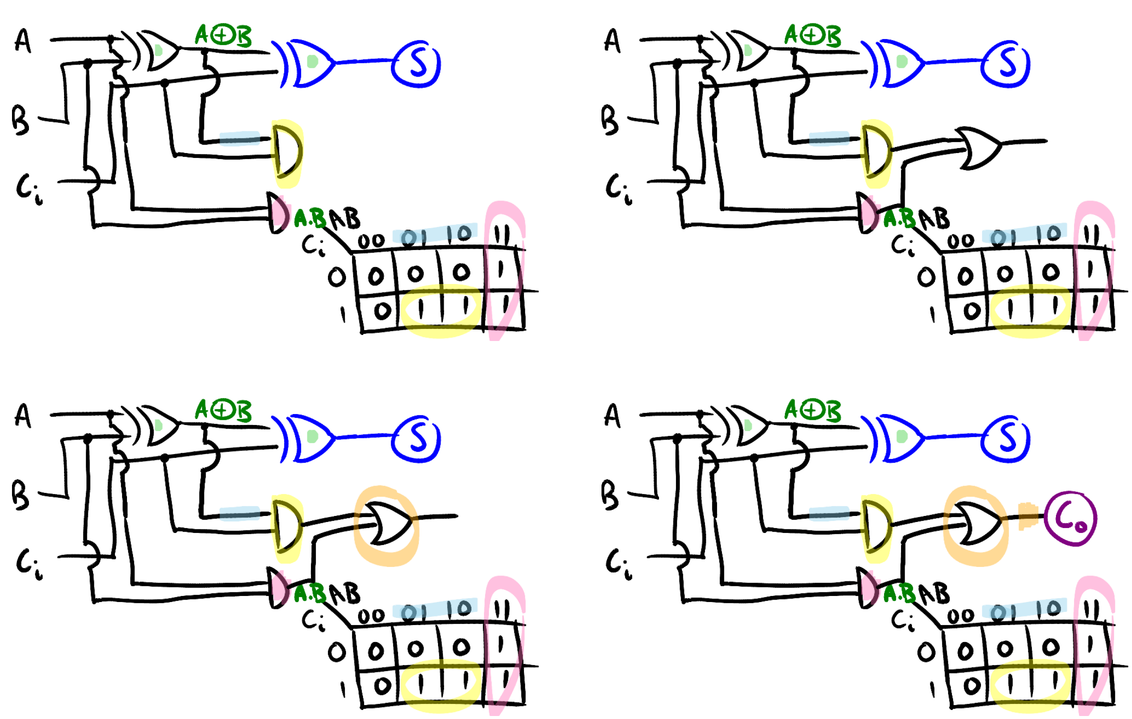


Figure 4-13: The full adder sketch as it is being drawn as described in Table 4.13. Notice that the system highlighted parts of the sketch when asking questions about them.

- the full system
- the system without dialogue capabilities
- the system without dialogue capabilities and without speech capabilities.

These comparisons will help provide some perspective. We will also ask the users how easy the systems are to use and how natural they are compared to working with traditional software and compared to working with another person.

Chapter 5

Related Work

5.1 Wizard-of-Oz studies

Wizard-of-Oz studies are common and have been conducted even in situations where the wizard simulates both pen and speech data [21, 22]. 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. Each of the participants in our study had unique design projects that they described, which were unknown before the user study.

5.2 Multimodal User Interfaces

Multimodal user interfaces originated with Bolt's "Put-That-There" system. Working in the domain of rescue mission planning, Bolt's system used pointing gestures to resolve designated keywords in the speech [5]. The field has gradually grown to include more interesting and complex non-verbal input.

QuickSet[20] is a collaborative multimodal system built on an agent-based architecture. 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. This differs from our goal of creating the most natural user interface possible. In contrast to our system where the user starts with a blank screen, QuickSet is a map-based system and the user starts with a map to refer to. Like our initial system, QuickSet uses a continuous speaker-independent speech recognition system.

All of these systems have benefitted from a series of empirical studies of multimodal communication. Oviatt et al. document users' multimodal integration patterns across speech and pen gestures in [22].

There are several other related projects[10, 20] that involve sketching and speech, but they are focused more on a command-based interaction with the user. In our system, speech augments the sketching; in other systems, the speech is necessary to the interaction.

Several existing systems allow users to make simple spoken commands to the system [9, 17]. We had many instances of users writing words and speaking them, which is very similar to the types of input that [17] handles. Kaiser et al. describe 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 [18]. As we have discussed they have made significant progress in learning new vocabulary. Our work will focus on learning the names and features of sketched objects. Also, our system will be interactive and run in real time; the work done by Kaiser et al. runs after a meeting has concluded which does not allow the system to interact with the users and ask the users questions.

5.3 Multimodal Dialogues

Focusing explicitly on managing multimodal dialogues, Johnston et al. describe MATCH in [16]. MATCH includes an 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 [23]. 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. Some recent work on multimodal reference resolution uses a greedy algorithm that uses linguistic and cognitive principles to efficiently resolve the references [6].

5.4 Our Previous Work

Our previous system, ASSIST [4], lets users sketch in a natural fashion and recognizes mechanical components (e.g., springs, pulleys, axles, etc.). Sketches can be drawn with any variety of pen-based input (e.g., tablet PC). ASSIST (see Figure 5-1) displays a "cleaned up" version of the user's sketch and interfaces with a simulation tool to show users their sketch in action.

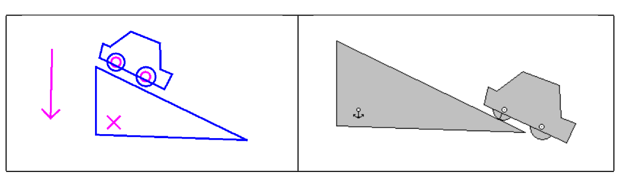


Figure 5-1: The left image shows the sketch in ASSIST. The right image shows the simulation.

Assistance [19] was a previous effort in our group to combine speech and sketching. It built on ASSIST by letting the user describe the behavior of the mechanical device with additional sketching and voice input.

Chapter 6

Contributions

Our goal is to develop a multimodal digital whiteboard that can understand and participate in a natural bidirectional conversation with a user who is engaged in a design task. The computer will be able to ask questions that resolve uncertainties and help the user refine their design. The contributions of this work including pursuing an open-ended, user driven multimodal dialogue. This differs in several ways from existing dialogue systems. First, the computer can respond and ask questions using both speech and sketching modalities. Second, in a typical dialogue system there are certain slots that need to be filled when making an airline reservation or asking about the weather. However, in a design scenario the goal of the user varies and there are no predetermined slots that the computer can ask questions to fill. The interaction is driven by the design the user is creating. Another contribution is that the dialogue is a key method that will help the computer resolve what the user is sketching and saying.

Building such a system will require solving some challenges along the way. These challenges include creating an effective representation for combing the speech and sketching inputs that includes information about the color of the sketching and properties of the speech. Other challenges will be learning new speech/sketching vocabulary and modifying the sketch in a generic extensible way. Generating the questions and determining how and when to ask them will also be a challenge in our work.

We conducted two user studies; the first study examined the relationship between speech and sketching. The more recent study investigated interactive multimodal dialogue. These studies will guide the construction of an interactive multimodal dialogue system. The dialogue study provided insight into the use of sketching, language, multimodal interactions and questions used in the dialogue. Importantly we found that speech phrases preceded sketching, contrary to earlier studies while the ordering of individual words and the corresponding sketch element matched the results from earlier research.

A central contribution of this work will be the use of multimodal dialogue to help the computer resolve the uncertainties that arise in multimodal interaction and to help the user by asking questions about their design. The goal is to move toward a situation where the user can have a natural conversation about a design with the computer like they would have with another person.

Appendix A

Proposed Schedule

1. Integrate Sketching and Speech (3 months) Jan-Mar 2007
2. Modifying the Sketch (1 month) Apr 2007
3. Learning (3 months) May-July 2007
4. Questions Asking / Dialogue (4 months) Aug - Nov 2007
5. Evaluation(2 months) Dec 2007 - Jan 2008
6. Writing (overlap with evaluation) (6 months) Dec 2007-May 2008

Bibliography

[1] Aaron Adler. Segmentation and alignment of speech and sketching in a design environment. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, February 2003.

[2] Aaron Adler and Randall Davis. Speech and sketching for multimodal design. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pages 214–216. ACM Press, 2004.

[3] James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. Towards conversational human-computer interaction. *AI Magazine*, 22(4):27–38, 2001.

[4] Christine Alvarado and Randall Davis. Resolving ambiguities to create a natural sketch based interface. In *Proceedings of IJCAI-2001*, August 2001.

[5] Richard A. Bolt. Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, pages 262–270, 1980.

[6] Joyce Y. Chai, Zahar Prasov, Joseph Blaim, and Rong Jin. Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In *IUI '05: Proceedings of the 10th international conference on intelligent user interfaces*, pages 43–50, New York, NY, USA, January 2005. ACM Press.

[7] P. R. Cohen, M. Johnston, D. R. McGee, S. L. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clowi. Quickset: Multimodal interaction for distributed applications. In *Proceedings of Mutimedia '97*, pages 31–40. ACM Press, 1997.

[8] Randall Davis. Sketch understanding in design: Overview of work at the MIT AI lab. *Sketch Understanding, Papers from the 2002 AAAI Spring Symposium*, pages 24–31, March 25-27 2002.

[9] David Demirdjian, Teresa Ko, and Trevor Darrell. Untethered gesture acquisition and recognition for virtual world manipulation. *Virtual Reality*, 8(4):222–230, September 2005.

[10] Kenneth Forbus, R. Ferguson, and J. Usher. Towards a computational model of sketching. In *Intelligent User Interfaces '01*, pages 77–83, 2001.

[11] James Glass, Eugene Weinstein, Scott Cyphers, Joseph Polifroni, Grace Chung, and Mikio Nakano. A framework for developing conversational user interface. In *Proceedings of the 4th International Conference on Computer-Aided Design of User Interfaces*, pages 354–365, January 2004.

[12] Tracy Hammond and Randall Davis. Tahuti: A geometrical sketch recognition system for UML class diagrams. *AAAI Spring Symposium on Sketch Understanding*, pages 59–68, March 25-27 2002.

[13] Tracy Hammond and Randall Davis. Ladder, a sketching language for user interface developers. *Elsevier, Computers and Graphics*, 28:518–532, 2005.

[14] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. In *Computer Speech and Language*, pages 49–67, 2002.

[15] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, and Ronald Rosenfeld. The SPHINX-II speech recognition system: an overview. *Computer Speech and Language*, 7(2):137–148, 1993.

[16] Michael Johnston, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 276–383, 2002.

[17] Edward C. Kaiser. Multimodal new vocabulary recognition through speech and handwriting in a whiteboard scheduling application. In *IUI '05: Proceedings of the 10th international conference on intelligent user interfaces*, pages 51–58, New York, NY, USA, January 2005. ACM Press.

[18] Edward C. Kaiser. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the Eighth International Conference on Multimodal Interfaces (ICMI 2006)*, Baff, Canada, November 2006.

[19] Michael Oltmans and Randall Davis. Naturally conveyed explanations of device behavior. In *Workshop on Perceptive User Interfaces*, 2001.

[20] Sharon Oviatt, Phil Cohen, Lizhong Wu, John Vergo, Lisbeth Duncan, Berhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. In *Human Computer Interaction*, pages 263–322, August 2000.

[21] Sharon Oviatt, Philip Cohen, Martin Fong, and Michael Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *Proceedings of the International Conference of Spoken Language Processing*, pages 1351–1354. University of Alberta, 1992.

[22] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Conference Proceedings on Human Factors in Computing Systems*, pages 415–422. ACM Press, 1997.

[23] Charles Rich and Candace Sidner. COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapter Interaction*, 8(3–4):315–350, 1998.

[24] Chang She. A natural interaction reasoning system for electronic circuit analysis in an educational setting. Master's thesis, MIT, June 2006.

[25] David G. Ullman, Stephen Wood, and David Craig. The importance of drawing in the mechanical design process. *Computers and Graphics*, 14(2):263–274, 1990.