Speech and Sketching for Multimodal Design

Aaron Adler and Randall Davis
MIT Computer Science and Artificial Intelligence Laboratory
77 Massachusetts Avenue, 32-239
Cambridge, MA 02139 USA
{cadlerun, davis}@csail.mit.edu

ABSTRACT

While sketches are commonly and effectively used in the early stages of design, some information is far more easily conveyed verbally than by sketching. In response, we have combined sketching with speech, enabling a more natural form of communication. We studied the behavior of people sketching and speaking, and from this derived a set of rules for segmenting and aligning the signals from both modalities. Once the inputs are aligned, we use both modalities in interpretation. The result is a more natural interface to our system.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*Natural language, Graphical user interfaces (GUI), Evaluation/methodology, Input devices and strategies (e.g., mouse, touchscreen), Interaction styles (e.g., commands, menus, forms, direct manipulation), User-centered design, Voice I/O*

General Terms

Performance, Design, Experimentation, Human Factors

Keywords

speech, sketch, multimodal interaction

1. INTRODUCTION

Sketches are commonly used in the early stages of design. Our previous system, ASSIST[2], 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 1) displays a "cleaned up" version of the user's sketch and interfaces with a simulation tool to show users their sketch in action.

Some parts of a mechanical system might be too difficult to express by sketching alone, but might be easy to describe verbally. In that case, adding speech recognition creates a more natural user interface. Our goal is to create a multimodal system where the user can have a natural conversation with the computer, of the sort a user might have with another person. We do not want the speech to be limited to simple, single word commands, like uttering "spring" while pointing. Rather, we want to allow the user to say whatever comes to mind and have the system gather as much as possible from the speech input [1].

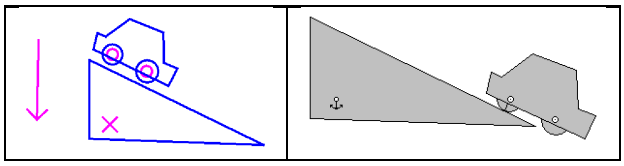

Figure 1: The left image shows the sketch in ASSIST. The right image shows the simulation.

We begin with an example that motivates our work, then describe how we collected data and created the set of rules for our system. Next, we describe how the speech and sketching components of the system are combined and conclude with related and future work.

2. MOTIVATING EXAMPLE

Newton's Cradle (see Figure 2) is 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 at the end of the pendulums 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 easy to create.

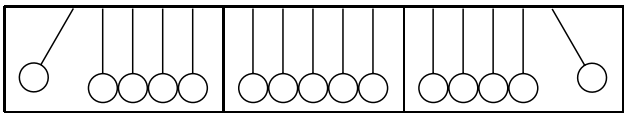

Figure 2: A sequence of images showing Newton's Cradle when one of the pendulums is pulled back and released.

3. OBTAINING SAMPLE DATA

To support natural speech, we conducted an empirical investigation of spoken descriptions of mechanical devices while the participant was drawing. We videotaped six outside participants while

Copyright is held by the author/owner.
IUI'04, January 13–16, 2004, Maderia, Funchal, Portugal.
ACM 1-58113-815-6/04/0001.

they sketched six mechanical systems at a whiteboard. They were given small hardcopy drawings of the systems and were told to draw them on the whiteboard, describing them as they did so. They were told to describe them as if they were talking to a small group of people, such as in a physics tutorial. The figures had marks to indicate identical components and identical distances. 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 using words we had chosen. The recordings from the participants were transcribed, and each speech and sketching action was time-stamped. This provided a basis for developing a set of approximately 50 rules that could segment and align the speech and sketching events.

4. SEGMENTING DATA

The data from the videos were analyzed by hand, segmenting it into individual speech events (roughly, phrases) and sketching events (drawing part of an object), and aligning corresponding events. From this analysis, we manually derived a set of rules that encapsulated the knowledge gathered. 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 (e.g., pauses are gaps of at least 0.8 seconds where there is no sketching or speech event), while others look for key words in the speech events. For example, words such as "and," "then," or "there are" were good indicators that the user started a new topic. In our analysis we noted that users never talked about one thing while sketching another.

The rules determine a set of times, or break points, that group together speech and sketching events that refer to the same objects. One rule indicates a possible break point when a speech utterance starts with a key word which is preceded by a pause. This might produce a group that included the speech phrase "that's suspended by springs on the bottom" and the three sketching events in which a spring is sketched.

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.

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.

There are three stages to the processing of the speech and sketching. The initial partitioning of both is done by the rule system. 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.

4.1 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 break points in the hand-generated segmentations. The computer-generated segmentation matched on 24 of these, and found 18 additional break points. The 18 additional break points were analyzed by hand and further classified as "incorrect," "inconsequential," or as resulting from "shallow knowledge." The "inconsequential" category includes break points that were immaterial to parsing, such as break points added at the beginning, prior to any speech or sketching events, and extra break points between some speech events at the end of the interaction (see Table 1). The "shallow knowledge" category contains additional break points that were placed between sketching events (see Table 2).

1a	"I'm puzzled as to how to indicate that"
2a	"equal size of"
2b	"the suspended balls"
3a	"and that it is not the same as"
3b	"the falling balls"

Table 1: Data from one of the participants exhibits how the speech we are working with is not grammatical. The hand segmentation placed all 5 events into the same group, however, the software placed the events into three groups by placing "inconsequential" break points between speech events 1a and 2a and between speech events 2b and 3a.

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

Table 2: Example of a "shallow knowledge" break point. The hand segmentation placed all 5 events into the same group, however the software placed the events into two groups by placing an extra break point between sketching events 1c and 2a. 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.

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.

5. SYSTEM OVERVIEW

Figure 3 shows screen shots of the working system.

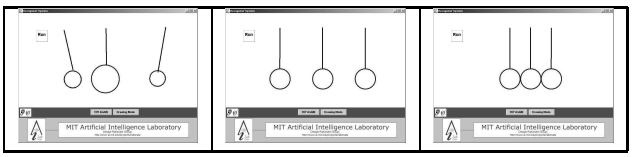

Figure 3: Three successive steps in our multimodal system. The first image shows the sketch before the user says anything. The second image shows the sketch after the user says "there are three identical equally spaced pendulums." The third image shows the sketch after the user says that the pendulums are touching.

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[4], 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.

ASSIST was modified so that the sketch interpretations were combined with the speech recognition data, possibly resulting in a modified sketch. 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.

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.

6. RELATED WORK

ASSISTANCE[6] 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. Our new system lets the users simultaneously talk in an unconstrained manner and sketch, which produces a more natural interaction.

QuickSet[7] 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 more 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 system, QuickSet uses a continuous speaker-independent speech recognition system.

AT&T Labs has developed MATCH[5], which provides a speech and pen interface to restaurant and subway information for New York City. This program uses a finite-state device and lets users make simple queries. 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.

There are several other related projects[3, 7] 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.

7. FUTURE WORK

Speech will allow the system to capture information that is not currently available with only the sketching interface. Speech is a rich input modality and more information, such as numerical references, can be extracted from it to aid in the disambiguation of the inputs. Future work will attempt to make it easier to add new objects and commands to the system. We also want to evaluate how people actually talk when presented with a working system of this type. Other input modalities, such as gesture, could also help disambiguate the sketches and correctly simulate the user's designs.

8. ACKNOWLEDGMENTS

This project was funded by MIT's Project Oxygen and iCampus.

9. REFERENCES

[1] A. Adler. Segmentation and Alignment of Speech and Sketching in a Design Environment. Master's Thesis, Massachusetts Institute of Technology, 2003.

[2] C. Alvarado and R. Davis. Resolving ambiguities to create a natural computer-based sketching environment. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 1365–1374, 2001.

[3] K. D. Forbus, R. W. Ferguson, and J. M. Usher. Towards a computational model of sketching. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, pages 77–83. ACM Press, 2001.

[4] T. J. Hazen, S. Seneff, and J. Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, 16:49–67, 2002.

[5] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383, 2002.

[6] M. Oltmans. Understanding Naturally Conveyed Explanations of Device Behavior. Master's Thesis, Massachusetts Institute of Technology, 2001.

[7] S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human Computer Interaction*, 15(4):263–322, 2000.