Creating a Multimodal Design Environment Using Speech and Sketching

Aaron Adler CADLERUN@AI.MIT.EDU

MIT Computer Science and Artificial Intelligence Laboratory, 200 Technology Square, Cambridge MA, 02139 USA

1. Introduction

Sketches are a commonly used tool in the early stages of design. Our previous system, ASSIST(Alvarado, 2000), lets users sketch in a natural fashion and recognizes mechanical systems. ASSIST also interfaces with a simulation tool to show users their sketch in action. ASSISTANCE(Oltmans, 2001) builds on ASSIST by allowing the user to use additional sketching and voice input to describe the behavior of the mechanical device after it has been sketched.

Some parts of a mechanical system might be too hard or too complicated to express by sketching alone, but adding speech recognition to the system 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, like a conversation that she would have with another person. Having simultaneous speech and sketching inputs provides the system with more information and allows the system to capture more details, including the rationale behind the design. We do not want the speech to be limited to simple, single word commands, like uttering "block" while pointing. Rather, we want to allow the user to say whatever comes to mind and have the system gather everything it can from the speech input (Adler, 2003).

2. Motivating Example

There is a system of pendulums called Newton's Cradle 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 on the other end of the system. Although this system seems simple enough to sketch, it is in fact nearly impossible to sketch so that it operates properly. The system works because the metal balls at the end of the pendulums just touch each other, and 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. This illustrates that speech can be used to clarify things that cannot be shown by sketching alone.

3. Obtaining Sample Data

Six subjects were videotaped while sketching six mechanical systems at a whiteboard. They were given small versions of the systems and told to enlarge them and describe the sketches as they would to a small group of people - like a physics tutorial. The figures had marks to indicate identical components and identical distances. The marks were provided to get an idea of how the subjects would describe identical or equally spaced objects without providing the subjects with any particular vocabulary. The recordings from the subjects were transcribed and each speech and sketching action was time-stamped. This provided a basis for developing a set of rules that could segment and align the speech and sketching events.

4. Segmenting Data

The data from the videos were analyzed by hand by creating a series of charts and graphs. A set of rules was created that encapsulated the knowledge that was gathered. Some rules relate objects that are a similar shape, others utilize the timing of the speech and sketching events, 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. Also, users never talked about one thing while sketching another. The rules build on each other and determine a set of times, or break points, that group the speech and sketching events. This process of segmenting and aligning the data allows us to use the inputs to disambiguate each other.

4.1 WATCH

JESS, an expert system shell for Java, was used as a rule engine for the system. A visualization tool, WATCH, was created to help analyze the output. WATCH allows the results of the rule system to be viewed in a timeline format, emphasizing the relationships between the rules.

4.2 Results

The results of running the rules were compared to the hand-generated results for 4 of the 24 data sets. The rules used a limited amount of information to segment and align the transcribed data. There were 29 break points in the hand-generated segmentations. The computer matched 24 of these. The computer found 19 additional break points, 18 of which were acceptable break points. 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 rules were kept general to avoid overfitting the data. Although the computer and hand-generated segmentation differed in many places, the segmentation is not the final step and it does not have to be perfect. As long as the segmentation provides an acceptable result, it should be sufficient to provide a good basis for further analysis and recognition.

5. System Overview

The system can be divided into two parts: speech and sketching. The two data sources then serve as the inputs to the rule system. Figure 1 shows screen shots of the working system.

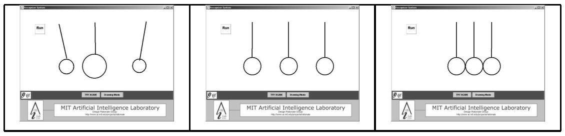
Recognizer System
Run
TRY AGAIN
Drawing Mode
MIT Artificial Intelligence Laboratory
Design Rationale Group
http://www.ai.mit.edu/projects/rationale

Figure 1. This figure shows three successive steps in the multi-modal 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.

5.1 Galaxy Speech System

The vocabulary and sentences from the transcribed videos, augmented with missing words, were used to create a speech recognizer for the system. The added words included numbers and plural words that were not in the transcripts. The Galaxy speech system, developed by the SLS group, is speaker independent and functions in a continuous recognition mode for our system. Both of these features are advantageous in our system. The speech recognition is accurate when the spoken words are included in the recognizer's vocabulary.

5.2 Modifying ASSIST

ASSIST was modified so that the sketch interpretations were combined with the speech recognition data. ASSIST was also modified to allow the system to change the resulting sketch. For example, for Newton's Cradle, functions were needed to equally space the pendulums and to make the pendulums identical. Changing the sketch required translating descriptions, such as "equally spaced," into a set of manipulation commands that were implemented in ASSIST.

The combined time-stamped data allowed the previously created rules to run on the real-time data. Some additional information was required in the system to recognize some other words from the speech input. For example, for Newton's Cradle, the word "pendulum" is very key. The system needs to know that a pendulum is a rod connected to a circular body so that it can find the corresponding shapes in the sketch. The combined system can respond to the sketching and the speech events and create identical and touching pendulums.

6. Future Work

Speech will allow the system to capture information that is not currently available with only the sketching interface. Adjusting various properties of the sketch would be easier with verbal interactions than with sketched input alone. Speech is a rich input modality and more information, such as numerical references, can be extracted from it to aid in disambiguation of the inputs. Future work will attempt to make it easier to add new objects and commands to the system. Other input modalities, such as gesture, could also help disambiguate the sketches and correctly simulate the user's ideas. Our next generation sketching system is blackboard-based(Hammond et al., 2002), which should allow the speech to be more fully integrated as a knowledge source on the blackboard.

7. Acknowledgements

I would like to thank my advisor Randy Davis for all his help. This project was funded by MIT's Project Oxygen.

References

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

Alvarado, C. (2000). A Natural Sketching Environment: Bringing the Computer into Early Stages of Mechanical Design. Master's Thesis, Massachusetts Institute of Technology.

Hammond, T., Sezgin, M., Veselova, O., Adler, A., Oltmans, M., Alvarado, C., & Hitchcock, R. (2002). Multi-domain sketch recognition. *MIT Student Oxygen Workshop*.

Oltmans, M. (2001). Understanding Naturally Conveyed Explanations of Device Behavior. Master's Thesis, Massachusetts Institute of Technology.