Using Speech and Sketching in a Design Environment

Aaron Adler & Randall Davis

What: While sketches are widely used in the early stages of design, not all aspects of a design are easily sketched; some are more easily described verbally. We will create a more natural interaction for the user by being able to handle both modalities, allowing users to sketch and talk simultaneously, while also enabling the computer to better understand the user's intentions.

Why: Our previous system, ASSIST[2], recognizes mechanical components (e.g., springs, pulleys, axles, etc.) that users sketch in a natural fashion with a pen-like input device. ASSIST displays a "cleaned up" version of the user's sketch and interfaces with a physics simulation tool to show an animated version of the sketch.

Newton's Cradle (see Figure 1) is a set of pendulums consisting of a row of metal balls on strings. When you pull back and release a number of balls on one end, after a nearly elastic collision, the same number of balls will move outward on the other end. Although this appears to be easily sketched, it is nearly impossible to draw so that it operates properly. The metal balls must just touch each other and the pendulums must be identical. If the user could simply say that "there are five identical, evenly spaced and touching pendulums," the device would be easy to create.

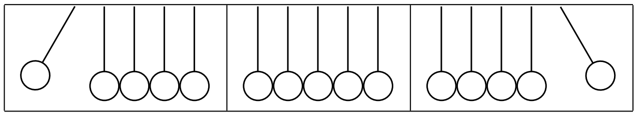

Figure 1: A sequence of images showing Newton's Cradle.

Previously, our group also built ASSISTANCE[5], which combined speech and sketching to allow users to describe the behavior of mechanical devices. ASSISTANCE builds on ASSIST by letting the user enter a second mode to provide the system with additional sketching and speech input. In contrast to ASSISTANCE, our new system lets the users simultaneously talk and sketch in an unconstrained manner, which allows for a more natural interaction.

There are several similar multimodal systems that incorporate speech and sketching. QuickSet[6] is a collaborative, multimodal, command-based system targeted toward improving efficiency in a military environment. The user can create and position items on an existing map using voice and pen-based gestures. QuickSet differs from our system because it is a command based system and users start with a map, which provides context, whereas our system uses natural speech and users start with an empty screen. AT&T Labs has developed MATCH[4], which provides a speech and pen interface to restaurant and subway information. Users can make simple queries using some multimodal dialogue capabilities. However, it uses command-based speech rather than natural speech, and it only has basic circling and pointing gestures for the graphical input modality not full sketching capabilities.

How: To support natural speech, we conducted an empirical investigation in which users were asked to sketch and verbally describe six mechanical devices at a whiteboard[1]. The participants were videotaped and the data was analyzed. The transcribed results revealed that the speech naturally used by people sketching and talking tends to be highly informal and unstructured. Despite this lack of standard grammatical structure, there turned out to be many clues about what the user is doing. Based on the data collected, we developed a set of approximately 50 rules to segment and align the speech and sketching inputs.

The rules segment the speech events and sketching events and then align corresponding events. Some rules group objects based on the timing between speech and sketching events (e.g., an overlapping speech event and sketching event are in the same group), while others look for objects that are the same shape (e.g. grouping consecutively drawn rectangles). Other rules looks for key words such as "and" or "there are" in the speech input. These key words were found to be good indicators that the user was starting a new topic, and disfluencies, such as "ahh" and "umm", were found to be indicators that the user was still thinking about the same topic.

The system has a grammar framework that recognizes certain nouns and adjectives. Nouns that the system can recognize include “pendulum,” while the adjectives include numbers and words such as “identical” and “touching.” The system needs to know the semantics of the sketch. For example, it needs to know that a pendulum is a rod connected to a circular body and it needs to know how to make the correct manipulations to the sketch to make such objects “touching.”

Progress: Two speech interfaces to the system have been implemented – one using ViaVoice and one using part of the SLS Galaxy speech system[3]. The Galaxy system is speaker independent, while the ViaVoice system can run independently in Windows on the same machine as the rest of the system. Both speech interfaces provide the time-stamped text of the speech utterance to the system. The rule system processes the incoming sketch data from ASSIST and speech events from the speech recognizer into groups of related events. The grammar framework then examines the events and possibly modifies the sketch. Figure 2 shows an example of the system working.

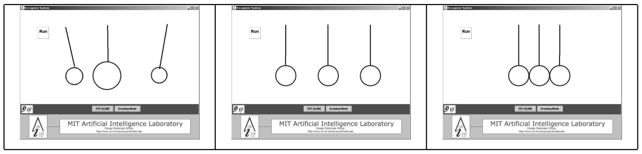
Recognizer System
Run
TRY AGAIN
Drawing Mode
MIT Artificial Intelligence Laboratory
Design Rationale Group:
http://www.ai.mit.edu/projects/rationale
Recognizer System
Run
TRY AGAIN
Drawing Mode
MIT Artificial Intelligence Laboratory
Recognizer System
Run
TRY AGAIN
Drawing Mode
MIT Artificial Intelligence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

Figure 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.

Future: Speech is a rich modality that captures information that is not available with sketching alone. Numerical references (e.g., “two,” “three”) can provide the system with information that would otherwise be difficult to convey. Adding new vocabulary to the system is difficult; we plan to investigate learning new vocabulary from the user by using clues that are in the speech utterances. For example, if the user talks about “three widgets” the system should know to look for three objects, even if it does not know what a “widget” is, and it should then attempt to learn what a “widget” is. As the system becomes more advanced, interacting with the user will help to clarify what the system does not understand. These steps will help to create an interaction that is as natural as possible.

Research Support: This work is supported in part by the MIT/Microsoft iCampus initiative and in part by MIT’s Project Oxygen.

References:

[1] 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.

[2] Christine Alvarado and Randall 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] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, 16:49–67, 2002.

[4] Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, 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 376–383, 2002.

[5] Michael Oltmans. Understanding Naturally Conveyed Explanations of Device Behavior. Master’s Thesis, Massachusetts Institute of Technology, 2001.

[6] 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. *Human Computer Interaction*, 15(4):263–322, 2000.