Mutual Disambiguation of Verbal and Sketching Inputs in a Design Environment

Aaron Adler, Randall Davis & Howard Shrobe

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: In any design environment, some things are more easily expressed verbally, while some are more easily expressed visually in a sketch. We would like to provide a design environment where the interaction is as natural as possible, and thus want to have both forms of input. In addition to making the interaction more natural, having speech recognition will allow mutual disambiguation between the sketch and the speech.

Motivation: Our previous sketching system [1] recognizes sketches of mechanical systems drawn with a pen-like stylus. However, we soon noticed that some things are difficult to express visually. For example, it is easy to say that we want three identical equally spaced objects, but far more difficult to draw this reliably. Also, speech input can supply a second source of information that can be used to disambiguate the sketch. This will create a natural and easy-to-use environment where the computer can assist the user without being obtrusive.

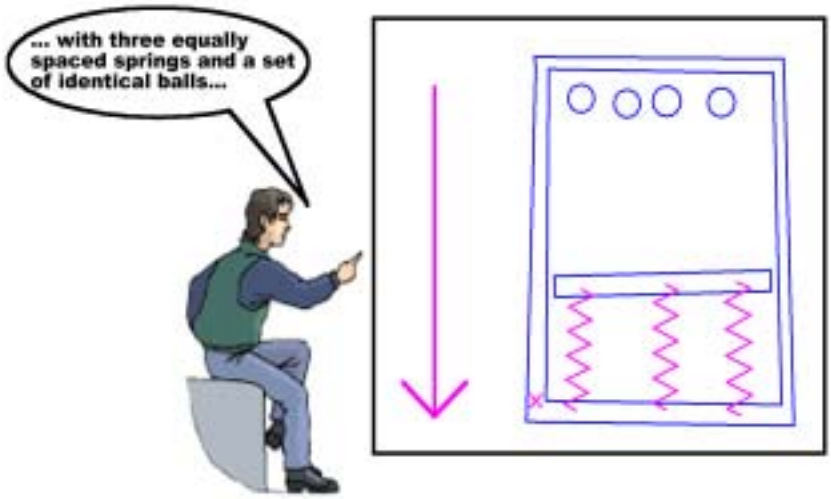

Figure 1: A user sketching and speaking to the system

Previous Work: Other work on sketching and multi-modal systems includes systems like QuickSet [5]. There has been research done on speech and sketching integration in QuickSet [3, 2]. Our work is different in that we want the interaction to be more than simply individual commands of the form ”place a bridge here.” We want to deal with speech that is not necessarily grammatical, which makes speech recognition a much harder task. The goal of our system is to obtain what information it can from the user’s natural speech and use that to disambiguate the sketch.

This work builds on our original sketching system [1] that has pen-based input. The original system also had a voice annotation mode to supplement sketched mechanical engineering systems [4]. The new system will allow sketching and speech inputs simultaneously.

Approach: We conducted an informal user study by videotaping people talking and sketching. The results were transcribed and broken down for analysis, revealing 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 that may help determine a variety of things about the interaction, including when the person sketching is switching topics. This is important because the sketching and speech need to be associated with each other, and time stamping alone is not sufficient. We also determined that there is no fixed temporal relation between sketching and vocalizations: people sometimes talked while sketching; sometimes spoke first, then drew; and sometimes drew before speaking. We observed that users occasionally give a general summary of what they are about to draw and then go back as they draw and describe the sketch in more detail. This information, the speech timing data, and even the pauses in speech should provide a wealth of information for the system to use.

We intend to collect a variety of such clues to infer as much about the structure of the speech and sketch as we can. With these clues we will create a simple rule based system to determine how well the rules can divide the inputs into groups of speech and sketch events.

We are currently working with the Speech and Language Systems group in LCS to develop a good model for the informal speech people seem to use. This speech will then be fed into the sketching system, which will combine the speech and sketching inputs to come up with the best possible interpretation of the sketch.

Impact: By allowing the user the ability to both talk and sketch, the system should be able to increase the accuracy of its interpretations. Multiple input modalities also allows the user to communicate more naturally with the computer. The hope is that the experience this creates for the user is one in which the user can interact in a very natural way with the system. It is also a long-term hope that the speech interaction will allow the system to clarify things that it doesn't understand with the user, which will provide the user with a natural and easy-to-use interface to the system.

Future Work: We would like the system to be able to interact with the user to seek clarifications when something is unclear. We would like the interaction with the computer to be as natural as possible and for the user to feel like they are interacting with another human. This could involve asking users intelligent questions at appropriate moments.

Research Support: This work is supported by MIT's Project Oxygen.

References:

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

[2] Michael Johnston. Unification-based multimodal parsing. In *COLING-ACL*, pages 624–630, 1998.

[3] Michael Johnston, Philip Cohen, David McGee, Sharon Oviatt, James Pittman, and Ira Smith. Unification-based multimodal integration. In Philip Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 281–288, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[4] Michael Oltmans. Understanding naturally conveyed explanations of device behavior. Master's thesis, MIT Artificial Intelligence Laboratory, Cambridge, MA, 2000.

[5] Sharon Oviatt and Philip Cohen. Mutimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3):45–53, 2000.