A Trimodal Dialogue Corpus: Speech, Gesture, and Sketching

Jacob Eisenstein JACOBE@CSAIL.MIT.EDU
Aaron Adler CADLERUN@CSAIL.MIT.EDU
Lisa Guttentag GUTTENTAG@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St., Cambridge MA, 02139 USA

1. Introduction

The development of perceptual user interfaces is central to Oxygen's mission of pervasive, human-centered computing. Based on speech, gesture, and sketching – rather than a keyboard and mouse – these new forms of interaction should more closely approximate human-human communication, which is natural and efficient.

To build computer interfaces around human-human interaction, human dialogues must be studied in more detail, from the perspective of ultimately building computer programs that can participate. Pure speech corpora, such as SWITCHBOARD (Godfrey et al., 1992), have greatly contributed towards the development of speech user interfaces. The next generation of multimodal user interfaces requires the development of corpora that include not only speech, but also non-verbal modes of communication, including gesture and sketching. Such corpora will help us understand how to interpret these modalities individually and how they interact with each other.

This paper describes the design, collection, and annotation of a new corpus of multimodal dialogues, as well as some potential applications. Developing a corpus on the order of SWITCHBOARD is a large-scale enterprise, consuming the complete attention of many full-time employees. In contrast, this corpus is a much smaller-scale effort, conducted mainly by students who hope to benefit from it in the course of our studies. We are not attempting to duplicate the size of such corpora, and consequently, some of the statistical techniques for leveraging corpora into language models (e.g., (Collins, 1997)) will not be applicable. We enumerate the benefits that we hope to attain from this corpus below.

1.1 Design Guidelines for Multimodal User Interfaces

One of the primary motivations for this research is to learn how non-verbal modalities are used in natural dialogues, so that we can design user interfaces that use these modalities in analogous ways. We hope to answer questions such as: What types of things are usually conveyed by gestures or by sketching? How are gesture, sketching, and speech interwoven in a coherent explanation? Most importantly, how can these findings be transformed into guidelines for the design of multimodal user interfaces?

For example, if it is found that many different speakers use the same handshape when describing a given object, then we would conclude that speakers usually tailor their handshape to the semantics of their speech. In this case, user interfaces may employ a complex and highly detailed handshape vocabulary as long as it is well suited for the semantics of the domain. However, if we find that handshapes are largely idiosyncratic, and that each speaker has preferred handshapes that are used without regard to semantics, then we would conclude that it is probably a bad idea for user interfaces to rely on detailed handshape vocabularies. These types of guidelines are relatively well understood for conventional graphical user interfaces, but at the moment, little is know about how to design multimodal user interfaces.

1.2 Test Bed for Multimodal Language Processing

There is a great deal of interest in multimodal language processing – extending NLP to other modalities such as prosody, gesture, and sketching (Quek et al., 2002). This research domain seeks to apply gestural cues to improve performance on a wide variety of natural language problems, such as topic and sentence segmentation, disfluency detection, and reference resolution.

However, the community currently lacks the standardized test bed corpora that have been so helpful for other areas of NLP. At the present, the development of a test corpus presents a significant barrier-to-entry for researchers who want to develop algorithms for multimodal language processing. Moreover, without standard corpora, it is impossible to compare competing systems.

1.3 Pilot for Future Studies

As mentioned above, the development of the large-scale corpora upon which data-driven NLP research has come to depend is a costly and time-consuming enterprise well be-

yond our current means. We hope that this study will serve as a pilot for designers of similar, more expensive, larger-scale corpora, who will be able to observe the successes and failures of our design.

To achieve this goal, we have taken steps to evaluate the quality of the corpus. We devoted one condition to the replication of earlier findings (McNeill, 1992), which includes findings regarding gesture type frequency, hierarchical organization of gestures, and the relationship between gestures and speech. We are compiling similar statistics from our study; if the results are similar on the conditions that are compatible, then the validity of both studies will be strengthened. In addition, we have conducted post-study surveys of all of our participants. These surveys are intended to give us an overall sense of how participants felt about the study, in particular whether they found anything about our setup to be distracting, whether they thought the task made sense, and whether they were able to guess the purpose of the study.

2. Procedure

Thirty college students and staff, aged 18-32, were chosen after responding to posters on the MIT campus. The data from two pairs of participants was not recorded correctly, leaving data from a total of 13 speaker-listener pairs including 15 females and 11 males. As determined by a pre-study questionnaire, English was not the first language of six of the participants. Of these, four were fluent in English, one was "almost fluent," and one spoke English "with effort."

McNeill and others have long advocated studying dialogues in which the speaker and listener already know each other (McNeill, 1992). This reduces inhibition, and eliminates a confound in which the speaker and listener gradually become less inhibited over time. Because of this, we recruited participants to sign up in pairs; 78% of participants described themselves as "close friends" or spouses of their partner; 20% as "friends", and 3% as "acquaintances".

We focused the dialogues on a specific topic, both to ensure that the data was meaningful and tractable, and to simulate the goal-directed collaboration that multimodal user interfaces try to attain. We chose the topic of mechanical design; the sketching and gesture metaphors in this domain are fairly obvious, as opposed to software design. In addition, the development of intelligent, multimodal interfaces for design is one of the long-term goals of our research group (Adler et al., 2004). Our participant pool was largely composed of people with some mechanical experience: 65% reported that they had taken a few physics or mechanics classes, and 27% reported that they used physics or mechanics frequently. This study may allow us to observe differences in explanation patterns between experts and non-experts. If significant differences are found, researchers interested in designing a multimodal user interface specifically for mechanical engineers would likely want to conduct a follow-up study with participants who are experts.

We also ran a condition outside of this domain, in which the focus of discussion was a "Tom and Jerry" cartoon. The cartoon domain has been studied extensively (McNeill, 1992); replicating previous results will help to validate our experimental design. In addition, we hope to be able to describe the differences in the types of gestures observed in the two domains.

The specific procedure for the study ran as follows. One participant was randomly selected to be the "speaker" and the other was the "listener." The speaker's job was to describe stories or mechanical devices to the listener. Prior to each description, the speaker either privately viewed a video of the relevant story or device or left the room and examined the actual device. Depending on the condition, the speaker was provided with either a whiteboard marker with which to create a sketch, a pre-printed visual aid, or no visual aids at all.

The listener's role was to understand these explanations and take a quiz later. The listener could ask questions of the speaker and was allowed to use the Tablet PC to take notes; a printed copy of these notes was provided during the quiz. A total of six topic conditions were run: four videos of simulations of mechanical devices, one physical mechanical device, and one cartoon. These topic conditions were balanced against the presentation conditions (marker, diagram, no visual aid), although the cartoon was always presented with no visual aid. Both the topic and presentation conditions were also balanced to eliminate ordering effects.

We found it necessary to limit the speaker to two minutes to view the video or object and three minutes to explain it. The majority of speakers used all of the time allotted. This suggests that we could have obtained more natural data by not limiting the explanation time. However, we found in pilot studies offering an unlimited amount of time led to problematic ordering effects, where participants devoted a long time to the early conditions, and then rushed through later conditions. With these time constraints, the total running time of the experiment was usually around 45 minutes.

2.1 Equipment and Setup

Our setup was guided by two principles. One goal was to keep the interaction as natural as possible, minimizing factors that would remind participants that they were in an experiment. Another goal was to achieve a very high quality recording of the participants, so that automated speech and gesture recognition could be performed. These goals

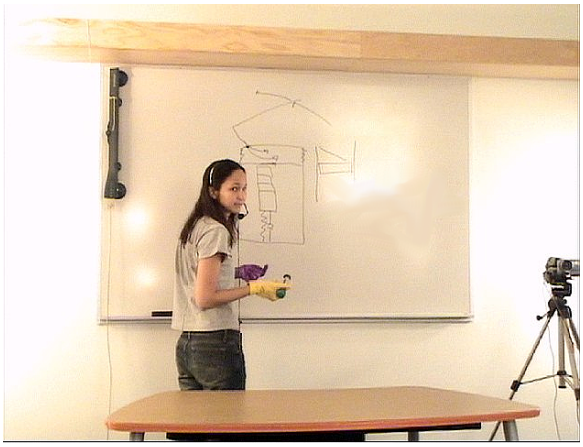

Figure 1. A participant drawing the Pez dispenser.

sometimes conflicted. For example, we considered using tracked gloves to obtain highly accurate 3D position information for each hand. In the end, however, we felt this would be too unnatural, and the resulting data might be badly skewed. Instead, we focused on vision-based techniques for hand tracking. In our pilot phase, we asked some participants to wear brightly colored gloves, and others to wear dark, long-sleeve shirts. Neither appeared to tip off participants as to the purpose of the study – in a post-study questionnaire, only one participant mentioned gestures or body language in his response. The gloves seemed more advantageous from a hand tracking standpoint, and moreover, the long-sleeve shirts were uncomfortably warm. This problem was compounded by the fact that the room was illuminated using 1380 watts of incandescent lights.

Separate cameras and headset microphones were used for the speaker and listener. We experimented with a lapel microphone for the speaker, but found the recording quality to be unpredictable. To ensure very tight audio-visual synchronization, we used "camcorder" style cameras with integrated audio recording components, rather than individual audio and video recording devices. Separate cameras and headset microphones were used for the speaker and listener. The camera output was encoded on the fly to MPEG format, using WinTV PVR-250 hardware encoding cards. This enabled us to use an under-powered computer (400 MHz, 128 MB ram) to perform the video capture.

A separate machine running Windows XP was used to capture the whiteboard drawing using the Mimio whiteboard capture device. In addition, the questionnaires for each participant were administered on separate Tablet PCs. The listener's tablet PC also recorded stroke data for any notes taken during the presentation, and the speaker's tablet PC was used to present the instructions. All four computers were coordinated using a client-server architecture written in Java and C# and administered from the Linux machine that was also responsible for the video encoding.

3. Current Status and Future Plans

In total, we now have roughly fourteen hours of video, along with time-stamped pen stroke information from both the Mimio and the Tablet PC. Our next step is to begin to annotate these videos to establish a ground truth test set of gesture and speech data. Annotations will include gesture type and composition, and the reference relationships between anaphoric pronouns in the speech and specific gestures. We will also attempt to automate some of this annotation, using vision and speech recognition. With a labeled corpus at our disposal, we can then begin to exploit the corpus, both as a source of design guidelines for multimodal user interfaces and as a test bed for multimodal natural language processing.

Acknowledgements

The authors would like to thank Christine Alvarado, Sonya Cates, Randall Davis, Tracy Hammond, Michael Oltmans, Metin Sezgin, Ron Wiken, and the kind people of the Language and Learning area in the fourth floor of the Gates tower for tolerating our longstanding occupation of their conference room.

References

Adler, A., Eisenstein, J., Oltmans, M., & Davis, R. (2004). The design studio of the future. *Making Pen-Based Interaction Intelligent and Natural (accepted, in press)*. AAAI Fall Symposium.

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. *Proceedings of ACL '97*.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research development. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (Vol. 1)* (pp. 517–520).

McNeill, D. (1992). *Hand and mind*. The University of Chicago Press.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., & Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 171–193.