SESSION 2: SPOKEN LANGUAGE SYSTEMS II

Wayne Ward, Chair

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Three papers in this session concern gathering spontaneous speech data. Two of the papers (AT&T and SRU) used a Wizard of Oz paradigm to interact with users via telephone. In the WOZ paradigm, subjects are led to believe that they are interacting with a machine when an experimenter is actually performing the machine's task. While this technique is used by many sites collecting spontaneous speech data, most of the systems use visual displays to present information to users. These two papers give some insight into the unique characteristics of the interactions when the data is presented over a telephone.

AT&T collected data for the Air Travel Information Service task. Since several other sites are collecting data for this same task, direct comparisons can be made between data gathered using visual ouput and that using speech output over a telephone. Since the ouput was verbal, particular attention had to be paid to the amount of information that was output. Information was summarized and compressed before being given to users. The AT&T system used the MIT Natural Language Understanding and database retrieval modules, so comparisons of data from these two sites is especially useful. The percentage of utterances that the system could not understand was significantly higher in the AT&T data. User behavior after errors was similar in the two systems, but more pronounced in the AT&T data. The rate of false starts and filled pauses was higher for AT&T than for other sites, but this may have been due to the fact that the experimenter rather than the user controlled the recording. When subjects explicitly press a button before speaking, they may compose utterances more carefully before they begin. It is clear that audio interaction paradigms present a unique set of problems for interactive Human-Machine communication.

SRU also collected speech data from subjects over a telephone. This study used a route planning task and focused on the difference between Human-Human interaction and Human-Machine interaction. One unique feature of this experiment was that users really accessed the system in order to get information rather than participating in an experiment using simulated scenarios. There were two conditions in the experiment, Human-Human and Human-Machine. There was no machine used in the H-M condition, but callers were induced to believe that they were talking to a machine by passing the experimenter's response through a "voice disguise unit" to make it sound as if it were produced by a machine. The only difference between the two conditions was the voice alteration. A standard opening phrase was spoken to each caller. Not enough data has been gathered thus far to allow reliable statistical analysis, but some interesting observations can be made. Many callers assumed that the machine knew when it was being addressed and made many aside remarks assuming that the machine would ignore them. Much of the dialog in the H-H condition concerned finding out about the capabilities of the service, while dialog in the H-M system was confined to getting route information. Significantly fewer words were spoken in each utterance in the H-M condition. SRU plans to conduct a larger version of this experiment in the near future. It should prove very useful to examine data from users actually performing a task (for real) and to contrast this with the behaviors seen in the simulated scenarios.

The third paper which reported on spontaneous speech data analyzed the performance of users and their subjective experiences with a Spoken Language System. The paper examines how speed/accuracy tradeoffs affect user perceptions of a system. Three versions of an ATIS system were used which represented different speed/accuracy tradeoffs. In a debriefing questionnaire, subjects answered several yes/no questions regarding system performance. Answers to questions regarding speed and accuracy were what intuition would suggest, given the tradeoff. One question, "Would you prefer this method to looking up the information in a book", seems to be more associated with overall user satisfaction. User responses to this question were not significantly different across systems.

The effect of user experience on recognition performance was also examined. In general, it was found that users who had a poor recognition rate on an initial scenario showed improved performance on a subsequent scenario. This improvement was correlated with a decrease in the perplexity of the utterances used. To some degree these subjects were able to adapt to the language model of the system. However, subjects with a relatively low initial error rate showed no improvement on the second scenario.

A third experiment investigated the effect of speaking style on recognition error. When recognition errors are made, subjects often try to help the system by changing their speaking style. As the authors point out, this degrades system performance since the system was not trained on this type of data. Instructing subjects not to hyperarticulate did not improve system performance. The authors suggest that training data should contain this type of speech so that there is a better match between the training and test conditions.

The fourth paper presents an overview of the research being done at LIMSI-CNRS. LIMSI has a very broad program of research covering Voice Dictation, Spoken Dialog, Natural Language Processing and Non Verbal and

Multimodal Communication. They are pursuing a very ambitious project using computer vision, natural language, knowledge representation, speech and gestures.