CSR DATA COLLECTION PILOT

Denise Danielson, Project Leader
Jared Bernstein, Principal Investigator

SRI International
Menlo Park, California 94025

PROJECT GOALS

The objective of the CSR Corpus Development is to collect and deliver a large corpus of continuous speech data to support DARPA research efforts in *continuous speech recognition* (CSR). The CSR corpus is intended to be task independent and to consist of speech that is similar to that which would be expected from eventual users of real world CSR systems. Toward these ends, the current pilot collection effort was designed to minimize controls on vocabulary, background noise, and microphones. CSR is also designed to incorporate spontaneous speech.

SRI was one of three sites to participate in a CSR pilot collection effort in 1991. The primary goals of the pilot effort were to:

- establish tools and procedures for CSR collection,
- experiment with various collection methods, including those for spontaneous speech,
- collect enough data for CSR tests in February 1992,
- estimate the cost and test the suitability of the data before proceeding with a full-scale CSR corpus.

RECENT RESULTS

Between October and December 1991, SRI completed its portion of the pilot CSR Corpus collection. The full design of the pilot CSR data collection is described in the paper "CSR Corpus Development" by George Doddington that appears in this volume. In SRI's portion of the task, a total of 13160 sentences from 36 speakers were collected, transcribed and delivered to NIST. This includes 1280 spontaneously produced sentences from 12 different speakers.

All data was collected simultaneously through two microphones. One microphone, a Sennheiser headset-mounted microphone, remained constant throughout all data collection. The second microphone varied across speakers and collection sessions. Depending on the type of data being collected, the second microphone was selected from one of two disjoint sets of microphones designated as either *training* or *test*. Thus, SRI's part of the CSR pilot corpus supports the development of "microphone independent" speech recognition systems. SRI's recordings take about 6.4 GBytes of storage for the two microphone channels, or about 3.2 GBytes per channel.

In each cell of the overall design, 50% of the subjects were male and 50% female. Half of the data in each condition includes verbalized punctuation, and half is without verbalized punctuation.

SRI has performed initial evaluations of the costs of collecting spontaneous speech data versus read speech data. In terms of subject time, spontaneous speech takes two to three times as long to produce as read speech. Because of the additional costs for transcription and materials preparation, the ratio in terms of data collection personnel time is closer to 6:1. In addition, since the task of creating and dictating news-style stories spontaneously requires specialized skills, cost per hour of subject time can be higher.

SRI is currently in the process of collecting an additional set of 8 test speakers. The current work includes experimentation with different materials and instructions for eliciting spontaneous speech. The goal of these experiments is to identify spontaneous collection methods that are relatively easy for the subject while yielding data that is truly spontaneous and useful for the system development efforts.

PLANS FOR THE COMING YEAR

- Work with DARPA and NIST to identify any changes necessary in the data collection procedures used during the pilot study.
- Proceed with a larger scale data collection effort.
- Perform more extensive analyses of the cost/benefit ratio of spontaneous speech data versus read speech data. Assuming these analyses show spontaneous speech to be cost effective in the long run, collect a larger proportion of spontaneous speech.
- Continue to experiment with various methods of eliciting spontaneous speech and identify those methods most appropriate for meeting the ultimate goals of the CSR corpus development project.