CSR CORPUS COLLECTION

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). SRI's current goal is the completion of Phase 2, Part 1 of the planned CSR Corpus. This consists of 86,000 sentences from 275 speakers, including 8000 spontaneous sentences from 40 journalists.

The Phase 2 Corpus collection task is a high volume data production task. SRI's major goal has been efficiency. Other goals include gathering data that is more representative of the *real world* by minimizing controls on vocabulary, microphones, background noise and speaker disfluencies, while improving data quality controls.

RECENT RESULTS

SRI began work on the current phase of CSR in September, 1992 and expects to complete delivery of this portion in June, 1993.

Data Production — As of 12 March 1993, SRI has collected the following portion of this CSR database:

Subject type	Required	Collected
Non-journalist long	25	22
Non-journalist short	210	125
Journalist	40	17

Subject Efficiency — SRI's first goal was to speed up subject interaction with the data collection software. Additional memory was added to the data collection systems, and data collection software made much faster, so that now the pace of the data collection process is directly controlled by the subject and no longer limited by the software. As a result, the average data collection pace has increased from 125 utts/hr to 200 utts/hr. For a typical short-term non-journalist subject collecting 190 read sentences, these changes and a faster paced orientation have reduced subject time from 120 minutes to 90 minutes. The shorter time requirement also makes it easier to attract and schedule subjects.

Process Efficiency — SRI has also been concerned with reducing the labor required to process speech data. A labor savings was realized by removing monitors from the data collection room. The data collection monitor now spends about 25 minutes instructing and observing while subjects collect their first few utterances, and then leaves the room. Two other changes have significanly improved labor efficiency. SRI has developed a new transcription tool that has led to a 15% to 20% reduction in transcription time and improved accuracy. We have also automated most pre-archival and archival steps.

Data Quality — SRI has incorporated NIST data quality software into its procedures. Sample files are collected at the start of each day on each data collection system. These files are run through the *wavmd* program, which runs a signal-to-noise (SNR) evaluation and other tests. Additional checks are performed on all files as they are collected to ensure that problems (e.g. dead microphone) are caught.

Labor Analysis — SRI is analyzing labor costs as we proceed with the current project to enable us to predict costs in the future, as well as to target specific tasks for efficiency improvements. A first round of labor analysis in January of this year identified transcription as one of the biggest labor costs. This has led to efforts to make the transcription task easier and more efficient. SRI continues to work with NIST and the CCCC to clarify transcription guidelines and implement changes recommended by CCCC. An analysis of recent project labor indicates that 10-15% of SRI's CSR project time has been spent on tasks in support of communication with NIST and various DARPA program committees.

PLANS FOR THE COMING YEAR

- Collect the remainder of the CSR Corpus.
- Work with NIST and CCCC to define goals and constraints for alternate microphones and environments.
- Work with NIST and the Data Quality Committee to further improve and automate quality tests of speech files.
- Work with NIST and the Data Quality Committee to define documentation requirements.
- Refine the spontaneous speech collection paradigm.