CSR DATA COLLECTION

Denise Danielson, Project Leader
Jared Bernstein, Principal Investigator

SRI International
Menlo Park, California 94025

PROJECT GOALS

The objective of the CSR Data Collection effort is to collect and deliver a large corpus of continuous speech data to support ARPA research efforts in continuous speech recognition (CSR).

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

RECENT RESULTS

SRI completed two separate CSR Data Collection projects in 1993:

Under the CSR Phase 2, Part 1 task, SRI delivered 86,000 utterances from 275 speakers, including 8000 spontaneous sentences from 40 journalists.

The CSR Development and Evaluation Spokes data collection task yielded 4435 development test utterances from 30 speakers and 4878 evaluation test utterances from a different set of 30 speakers. The development test data covered eight different spoke conditions, each of which had its own distinct combination of subject, prompt text, microphone and recording environment requirements. Similarly, the evaluation test data covered nine different spoke conditions and two hub conditions, each of which had its own unique requirements.

Efficiency — As a result of data collection software improvements the average data collection pace went from 125 utts/hr during the CSR Pilot Corpus Collection to 200 utts/hr for CSR Phase 2, Part 1. For a short-term non-journalist subject collecting 190 read sentences, these changes and a faster paced orientation reduced subject time from 120 minutes to 90 minutes. A number of changes also yielded savings in terms of SRI labor. The data collection supervisor now spends only about 25 minutes instructing and observing while subjects collect their first few utterances, and then leaves the room. Development of a new transcription tool led to a 15% to 20% reduction in transcription time and improved accuracy.

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. More comprehensive data quality checks and calibration procedures were established as part of the CSR Dev and Eval Spokes collection.

Documentation— SRI delivered complete documentation of data collection procedures, subject instructions, subject profiles, microphones and data collection environments for both the CSR Phase 2, Part 1 Corpus and the CSR Dev and Eval Spokes data. SRI views this as a first step in establishing documentation standards for speech corpora.

PLANS FOR THE COMING YEAR

SRI is currently between projects for CSR. We hope to participate in the CSR Phase 2, Part 2 data collection effort beginning in spring 1994. As part of this anticipated project, SRI plans to:

- Provide the CCCC with input from the data collection perspective.
- Test and implement the new data collection system currently being developed at MIT.
- Work with the CCCC, LDC and NIST to further improve and automate quality tests of speech files.
- Implement improved transcription conventions.
- Work with the LDC and CCCC to further define and improve documentation standards.