SESSION 11: CONTINUOUS SPEECH RECOGNITION AND EVALUATION I*

Clifford J. Weinstein, Chair

Lincoln Laboratory, M.I.T.
Lexington, MA 02173-9108

This was the first of two companion sessions which marked an important transition in the continuous speech recognition (CSR) component of the DARPA Spoken Language Program. Since 1987, DARPA CSR systems have been developed and evaluated on the Resource Management (RM) CSR corpus, which has become a *de facto* standard for comparison of speech recognizers, widely accepted and used both within and outside the DARPA research community, and internationally. The RM corpus has served the community well, but due to its limitations (e.g., 1000-word vocabulary, restricted task domain), it has become clear that more challenging tasks were needed to drive CSR research. The papers in this session reported on the design and preliminary development and analysis of a new large-vocabulary CSR corpus for the 90s, which is aimed at meeting the needs of CSR research, and at serving in a complementary role to the corpora being collected in the interactive Air Travel Information System (ATIS) domain.

The first major component of the new speech corpus is based on a very large Wall Street Journal (WSJ) text corpus, and supports recognition vocabularies of 5,000 and 20,000 words and higher. Actual corpus collection began in October 1991, and by the time of this February 1992 workshop, not only had a pilot WSJ corpus of 80 hours of speech (under varied conditions) been collected and distributed, but also a dry run evaluation of a number of CSR systems had been conducted; the papers in Session 12 describe those systems, tests, and results. This is a very significant accomplishment on such a short time scale, which was achieved through a multi-site effort featuring strong and effective cooperation in the context of multiple, sometimes conflicting goals, and some outstanding individual efforts.

The first paper in Session 11, presented by Janet Baker and Doug Paul, reported on the design of the WSJ-based CSR corpus, and of the pilot portion of the WSJ corpus. The paper described the efforts of the CSR Corpus Committee (chaired by Janet Baker, and including representatives from all the participating sites) in working out a design to meet multiple research goals. Key elements of the design which were outlined include: variable vocabulary sizes and perplexities; variable amounts of data per speaker to support speaker-dependent, speaker-adaptive, and speaker independent recognition paradigms; speech collected both with and without verbalized punctuation; simultaneous close-talking microphones and multiple secondary microphones; and numerous additional features. The paper described the text-processing steps performed at Lincoln on the original WSJ text CD-ROM to produce the texts and the language models which were delivered to the recording and testing sites. The paper summarized the materials delivered to the collecting and testing sites, which included: prompting texts for recording; truth texts for training, recognition, and scoring; a 33,000-word dictionary, provided by Dragon Systems, to cover training and test sets; a specification of training and test sets; and baseline language models for research and cross-site comparative testing.

The next paper, presented by George Doddington, described the work of the continuous speech corpus coordinating committee (CCCC), which was formed in October 1991 to oversee the actual development of the corpus and the subsequent test and evaluation. Doddington, chair of the CCCC, described the efforts of the committee and acknowledged the specific efforts of the individuals and sites involved. The February 1992 goals for the pilot corpus were met and exceeded. The collecting sites (MIT, SRI, and TI) had an initial target of about 45 hours, but actually collected a total of about 80 hours of read speech, while SRI collected a substantial corpus of spontaneous speech in the general WSJ domain. At the close of his talk, Doddington outlined several corpus issues to be reviewed, based on the pilot corpus experience, in proceeding with the full corpus collection and evaluation. These issues included: verbalized punctuation (VP) vs non-verbalized punctuation (NVP); natural vs preprocessed prompting text; mix of spontaneous vs read speech; and multiplicity of evaluation conditions. The discussion of these issues was tabled until the end of Session 12.

*This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government.

The next paper, presented by Jim Glass, described collection and analyses of WSJ-CSR data at MIT. Close attention was paid to development of an easy-to-use computer interface which enabled efficient data collection with minimal supervision of users. This interface was quite successful, and is currently being used by SRI and by NIST, as well as MIT. MIT collected 33 hours of speech and delivered the data to other sites on in-house-produced CD-ROM-compatible WORM disks. Experiments were described to investigate the effects of the text preprocessing performed on the WSJ text, by comparing sentences spoken with and without text preprocessing. The results showed a substantial variation in the readings of the unprocessed text, relative to the processed prompting text.

Finally, Jared Bernstein described experiments at SRI International in collection of spontaneous speech for the CSR corpus. The methods proved sufficient to collect fluent spontaneous speech, although at significantly greater cost and with significantly greater variation than the read speech. In particular, good subjects were hard to find. The paper described the procedures and quantified the effects. A number of interesting and entertaining samples of spontaneous renditions of news articles were played.

A few of the items raised in discussion were:

1. How about recording a paragraph at a time?

   D. Paul: The recording of sentence-at-a-time within paragraphs was more convenient for collection and for current research recognizers which are oriented to one sentence-per-file.

   V. Zue: Subjects were presented with highlighted paragraphs, and moved quickly from sentence to sentence.

2. Short discussion on verbalized punctuation.

   Janet Baker: People who dictate, do punctuate. Would like to see some data collected in a realistic dictation task (e.g., tell a reporter that a story due tomorrow morning must be dictated).

3. **Roger Moore**: Can Europeans get this data and tools?

   General encouragement on sharing the data, and on cooperative efforts.

4. **P. Price**: How about a description of the TI collection effort?

   R. Rajasekaran: TI used automatic endpointing rather than push-to-talk; most subjects didn't like verbalized punctuation.

Rest of discussion tabled till after dinner.

An unscheduled video presentation following Bernstein's talk entertained the group with Victor Borge's rendition of the sounds of punctuation. This video, courtesy of Rich Schwartz of BBN, was an excellent lead-in to the dinner break.