DARPA FEBRUARY 1992 PILOT CORPUS CSR "DRY RUN" BENCHMARK TEST RESULTS

David S. Pallett

National Institute of Standards and Technology
Building 225, Room A216
Gaithersburg, MD 20899

1 Introduction

Continuous speech recognition research activities within the DARPA Spoken Language community have, within the past several years, been focussed on the Resource Management (RM) and Air Travel Information System (ATIS) corpora. Within the past year, plans have been developed for a large, multi-component "general-purpose English, large vocabulary, natural language, high perplexity corpus" known as the DARPA [Wall Street Journal-based] Continuous speech Recognition (CSR) Corpus [1]. Doug Paul, of MIT Lincoln Laboratory (MIT/LL), and Janet Baker, of Dragon Systems, are responsible for many of the details of these plans. This corpus is intended to supplant the RM corpora and to supplement the ATIS corpora as resources for the DARPA speech recognition research community.

Plans to coordinate the design and collection of the CSR Corpus have, since October 1991, been coordinated by the DARPA CSR Corpus Coordinating Committee (CCCC), chaired by George Doddington, following discussions held by an earlier group [2].

In a meeting held at MIT Laboratory for Computer Science (MIT/LCS) in August of 1991, plans were developed for an initial "Pilot Corpus" comprising approximately 40 hours of recorded speech material, which was to be made available within the DARPA community in adequate time in order to permit reporting preliminary or "dry run" benchmark tests at the February 1992 meeting.

Following that meeting, NIST, acting as a DARPA "agent", contracted with Texas Instruments and SRI International (SRI) [3] for collection of the Pilot Corpus and the spoken language group at MIT/LCS also agreed to collect a substantial amount of material for the Pilot Corpus [4]. NIST prepared the material for production on recordable CD-ROM media (at MIT/LCS) and screened the associated transcriptions for conformance to standards. The group at SRI was the only group that collected "spontaneous dictation" in addition to the "read speech" comprising the bulk of the Pilot CSR Corpus.

More than 80 hours of material (per microphone channel) had been collected and distributed to several DARPA contractors. This material included Speaker-Dependent, Longitudinal Speaker- Dependent, and Speaker-Independent training components as well as specifically designated Development Test sets.

On January 17th (approximately one month before the Speech and Natural Language Workshop), two CD-ROMs containing a selected portion of the Pilot Corpus's Evaluation Test set were were distributed by NIST to four sites: CMU, Dragon Systems, MIT/LL, and SRI International. These sites had indicated interest in participating in the initial "dry run" benchmark test associated with the CSR Pilot Corpus.

2 Benchmark Test Material

The selected portion of the Evaluation Test Set that was distributed for use in the "dry run" benchmark tests, like the training material, included three major components: (1) Longitudinal-Speaker-Dependent speech recognition system test material, for use with the 3 speakers for which approximately 2400 CSR WSJ sentence utterances, per speaker, were available for speaker-dependent system training, (2) Speaker-Dependent system test material, for use with a set of 12 speakers for which 600 sentence utterances were available for speaker-dependent system training, and (3) Speaker-Independent system test material, with 10 speakers. The data was further broken down into verbalized-punctuation (VP) and non-verbalized-punctuation (NVP) and 5,000- vs. 20,000-word vocabularies.

For the purposes of speaker-independent system development, a specific set of approximately 7200 utterances obtained from an independent set of 84 speakers included in the training portion of the corpus had been designated with the concurrence of the CCCC.

The test material included material from SRI, MIT/LCS, TI and NIST (for one Speaker Independent

subject). Approximately 50% was from male speakers, and 50% female.

As noted elsewhere [1-2], the training and test material was selected with reference to predefined 5,000-word and 20,000-word lexicons, but with a controlled percentage of out-of-vocabulary (OOV) items. NIST's analysis of the test sets indicate that the actual occurence of OOV items in the 5,000 word SI test material is approximately 1.4% to 2.0%, and for the 20,000 word SI test material, it is 2.0% to 2.5%. In contrast, for the SI spontaneous test set, the incidence of OOV items with respect to the 5,000 word closed language model is 13.2% to 15.6%, and 5.3% to 5.6% with respect to the 20,000 word language model.

For this Pilot CSR Corpus, data was collected with both "primary" and "secondary" microphones. In every case, the primary microphone was a member of the Sennheiser close-talking, supra-aural headset- mounted, noise cancelling family (e.g., HMD-414, HMD-410). However, the microphones used as the secondary microphone were varied, and included boundary effect surface-mounted microphones such as the Crown PCC-160, Crown PZM-6FS, and Shure SM91).

3 Benchmark Test Protocols

The CCCC had agreed that, insofar as very little time had been allocated for system development and use of the Training and Development Test material, the contractor's results would not be reported to NIST until February 17th, less than one week prior to the Speech and Natural Language Workshop. It was also agreed that existing scoring software would be used as well as previously established procedures for scoring and reporting speech recogntion benchmark tests.

The four sites (CMU [5], Dragon Systems [6], MIT/LL [7] and SRI [8]) provided NIST with a total of 22 sets of results for a number of test sets and system configurations. The number of test set results provided by individual contractors ranged from 1 to 10.

NIST reported scores back to the contractors on February 19th. Subsequently, small discrepancies (typically less than one percent in the individual speakers' scores) were noted between the scores that had been determined at the individual sites and NIST's scores. Some of these discrepancies were due to a problem in handling the occurence of left parenthesis characters, "(", in the hypothesis strings in NIST's scoring program, and these differences were resolved after the Workshop. Consequently, there may be small unresolved differences between scores reported in this paper and others in this Proceedings.

4 "Best" Dry Run Evaluation Test Benchmark Test Results

The DARPA Spoken Language community's efforts to collect, annotate, process, and distribute the Pilot Corpus were challenging and highly stressful. It was generally agreed that there was insufficient time for system development between release of the training data and reporting "dry run" results, and that the systems for which results could be reported at the meeting represented only preliminary efforts.

Papers presented at the meeting typically include comments such as: "The training paradigm outlined... in the description...has only recently been fully implemented..." and "there has not yet been any opportunity for parameter optimization."[6], or "The tests... reported here are little more than pilot tests for debugging purposes and no strong conclusions can be drawn."[7] and "Our strategy was to implement a system as quickly as possible in order to meet the tight CSR deadline.[8]"

In view of these comments, and because comparisons across sites would be inconclusive, only a selected subset of results reported at the meeting are included in this paper. Several of the participants suggested that it would be acceptable to cite the "best" scores, based on lowest word error rate in a given test subset, and to do so without attribution to any specific system.

The "dry run" test results included in this paper (Table 1), are restricted to those selected "best" reported scores, and are presented without attribution to specific systems or sites. References 5 to 8 may contain additional information defining the context of these scores, or contain complementary findings based on experience with the development test sets. Caution should be exercised in interpreting these results as a valid indicator of the state-of-the-art, because of the short time for system development and debugging (as noted above).

5 Discussion

The initial "dry run" test results indicate general trends. Many of these trends would seem to be obvious, but are noted because one of the purposes of the CSR Pilot Corpus and the "dry run" was to verify the community's expectations with respect to challenges inherant in large-vocabulary continuous speech recognition, and to guage the relative significance of many factors.

- Results for the test sets selected from a smaller vocabulary (5,000 vs. 20,000 words) have lower error rates (e.g., for the longitudinal speaker dependent

speakers, for VP, 6.7% word error for the 5,000 word test subset vs. 10.6% for the 20,000 word test subset.

- Results for better-trained speaker dependent systems are better than for less-well-trained speaker-dependent systems (e.g., 6.7% word error for the test subset for the longitudinal speaker dependent speakers (5,000 word VP) vs. 14.7% for the speaker dependent speakers, for which one-fourth as much training material was available for each speaker).
- Results for speaker-independent systems have higher error rates than for speaker-dependent systems (e.g., 16.6% word error for the Speaker Independent subset (5,000-word VP) vs. 12.9% for the corresponding Speaker Dependent subset and 6.7% for the Longitudinal Speaker Independent subset).
- Results for the VP test subsets in general have lower error rates than for the NVP test subsets.
- Comparison of spontaneous vs. "read spontaneous" data indicates that the read spontaneous has lower error rates (as had been noted with earlier ATIS0 data).

The results for the challenging "spontaneous" and "read spontaneous" speech test subsets are based on only one sites's processing of the test data.

Only one site [8] reported results using both the primary and the secondary microphone(s) for the 5,000 word speaker Independent VP subset, reporting 16.6% word error for the primary microphone and 26.0% for the secondary microphone data. The incremental degradation in performance was regarded by the developers as less than might have been expected and "noteworthy" [9], particularly in view of the fact that the broadband signal to noise ratio for the secondary microphone data was typically 20 to 30 dB less than that for the primary microphone data.

Substantial variability in the rank-ordering of individual speakers can be noted across systems for those data subsets for which more than one site's or system's responses were reported. Analysis of this data suggests that some systems had greater variances across the speaker population than others, perhaps because of inadequate time to develop robust speaker-independent models.

NIST's measurements of the broadband S/N ratios for the primary microphone data from MIT, SRI, and TI range from 40 to 48 dB with values for the secondary microphone some 20 to 30 dB less than that. Histograms showing the distribution of levels for these files reveal evidence of gain changes between sessions and of the use of different secondary microphones for different data collection sessions (i.e., for the adaptation sentences vs. the read Wall Street Journal sessions).

Although reservations have been expressed by the participants in this initial "dry run" test, it should be noted that the results are highly encouraging in many ways. As the participants noted, "The successful application of... to the WSJ-CSR task demonstrates the utility of..." and "We have also demonstrated the utility of... in the context of a much larger task".[5] and "It is encouraging that... given there has not yet been any opportunity for parameter optimization."[6], and "The results, however, show promise and will require more rigorous testing."[7] and "This is a preliminary report demonstrating that... was ported from a 1000-word task (ATIS) to a large vocabulary task (5000-word) task... in three man weeks."[8]

Based on these observations, and on the experience gained in designing, collecting, and distributing the DARPA Pilot CSR Corpus, and in rapidly adapting existing technology to the new domain, there is good reason to look forward to the results of future benchmark tests.

6 Acknowledgements

At NIST, John Garofolo has been the individual responsible for coordinating and screening much of the CSR data collected at MIT/LCS, SRI and TI. Brett Tjaden assisted in preparation of the master tapes for CD-ROM production at MIT/LCS. Jon Fiscus adapted the NIST speech recognition scoring software for scoring the test results and implemented the software in preparing the official results.

7 References

1. Paul, D.B. and Baker, J.M., "The Design for the Wall Street Journal-based CSR Corpus", in Proc. Speech and Natural Language workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
2. Doddington, G.D., "CSR Corpus Development" in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
3. Phillips, M. et al., "Collection and Analyses of WSJ-CSR Data at MIT", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
4. Bernstein, J. and Danielson, D., "Spontaneous Speech Collection for the CSR Corpus", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.

5. Alleva, F. et al., "Applying SPHINX-II to the DARPA Wall Street Journal CSR Task", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
6. Baker, J. et al., "Large Vocabulary Recognition of Wall Street Journal sentences at Dragon Systems", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
7. Paul, D.B., "The Lincoln Large-Vocabulary HMM CSR", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
8. Murveit, H., Butzberger, J. and Weintraub, M., "Performance of SRI's DECIPHER Speech Recognition system on DARPA's CSR Task", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
9. Murveit, H., personal communication to D.S. Pallett, Mar.6, 1992.

		Word Err.	Sent. Err.
1(a) Longitudinal Speaker Dependent			
"5,000 Word"	NVP	10.6	75.6
	VP	6.7	55.6
"20,000 Word"	NVP	17.4	82.2
	VP	14.7	86.7
1(b) Speaker Dependent			
"5,000 Word"	VP	12.9	73.3
1(c) Speaker Independent			
"5,000 Word"	NVP	17.1	77.5
	VP	16.6	80.0
"20,000 Word"	NVP	37.9	94.5
	VP	32.8	93.4
1(d) "Spontaneous" Speaker Independent			
"Spontaneous"	NVP	55.8	100.0
	VP	45.5	97.0
"Read Spontaneous"	NVP	50.2	99.0
	VP	41.3	98.0

Table 1: "Best" Reported Scores, Selected on Word Error Percentage