DARPA Resource Management Benchmark Test Results June 1990

D. S. Pallett, J. G. Fiscus, and J. S. Garofolo

Room A 216 Technology Building
National Institute of Standards and Technology (NIST)
Gaithersburg, MD 20899

Introduction

The June 1990 DARPA Resource Management Benchmark Test makes use of the first of several test sets provided with the Extended Resource Management Speaker-Dependent Corpus (RM2) [1]. The corpus was designed as a speaker-dependent extension to the Resource Management (RM1) Corpus [2], consisting of (only) four speakers, but with a large number (2400) of sentence utterances for each of these speakers for system training purposes. The corpus was produced on CD-ROM by NIST in April 1990, and distributed to DARPA contractors. Results have been reported to NIST for both speaker-dependent and speaker-independent systems, and the results of NIST scoring and preliminary analysis of these data are included in this paper. In addition to the June 1990 (RM2) test set results, some sites also reported the results of tests of new algorithms on test sets that have been used in previous results ("test-retest" results), or for new (first-time) use of previous test sets, cr for new systems in development. Those results are also tabulated.

Test Protocol

Test results were submitted to NIST for scoring by the same "standard" scoring software used in previous tests [3] and contained on the CD-ROM version of the RM2 corpus. Minor modifications had to be made in order to accommodate the larger volume of test data. (For each of the four speakers, there were a total of 120 sentence utterances, so that the test consisted of a total of 480 sentence utterances, in contrast to the test set size of 300 sentence utterances used in previous tests.) Scoring options were not changed from previous tests.

Tabulated Results

Table 1 presents results of NIST scoring of the June 1990 RM2 Test Set results received by NIST as of June 21, 1990.

For speaker-dependent systems, results are presented for systems from BBN and MIT/LL [4] for two conditions of training: the set of 600 sentence texts used in previous (e.g., RM1 corpus) tests, and another condition making use of an additional 1800 sentence utterances for each speaker, for a total of 2400 training utterances. For speaker-independent systems, results were reported from AT&T [5], BBN [6], CMU [7], MIT/LL [4], SRI [8] and SSI [9]. Most sites made use of the 109-speaker system training condition used for previous tests and reported results on the RM2 test set. BBN's Speaker Independent and Speaker Adaptive results [6] were reported for the February 1989 Test sets, and are tabulated in Table 2. SRI also reported results for the case of having used the 12 speaker (7200 sentence utterance) training material from the speaker-dependent RM1 corpus in addition to the 109 speaker (3990 sentence utterance) speaker independent system training set, for a total of 11,190 sentence utterances for system training.

Table 2 presents results of NIST scoring of other results reported by several sites on test sets other than the June 1990 (RM2) Test Set. In some cases (e.g., some of the "test-retest" cases) the results may reflect the benefits of having used these test sets for retest purposes more than one time.

Significance Test Results

NIST has implemented some of the significance tests [3] contained on the RM series of CD-ROMs for some of the data sent for these tests. In general these tests serve to indicate that the differences in measured performance between many of these systems are small -- certainly for systems that are similarly trained and/or share similar algorithmic approaches to speech recognition.

As a case in point, consider the sentence-level McNemar test results shown in Table 3, comparing the BBN and MIT/LL speaker dependent systems, when using the word-pair grammar. For the two systems that were trained on 2400 sentence utterances, the BBN system had 426 (out of 480) sentences correct, and the MIT/LL system had 427 correct. In comparing these systems with the McNemar test, there are subsets of 399 responses that were identically correct, and 26 identically incorrect. The two systems differed in the number of unique errors by only one sentence (i.e., 27 vs. 28). The significance test obviously results in a "same" judgement. A similar comparison shows that the two systems trained on 600 sentence utterances yield a "same" judgement. However, comparisons involving differently-trained systems do result in significant performance differences -- both within site, and across sites.

Table 4 shows the results of implementation of the sentence-level McNemar test for speaker-independent systems trained on the 109 speaker/3990 sentence utterance training set, using the word-pair grammar, for the RM2 test set.

For the no-grammar case for the speaker-independent systems, the sentence-level McNemar test indicates that the performance differences between these systems are not significant. However, when implementing the word-level matched-pair sentence-segment word error (MAPSSWE) test, the CMU system has significantly better performance than other systems in this category.

Note that the data for the SRI system trained on 11,190 sentence utterances are not included in these comparisons, since the comparisons are limited to systems trained on 3990 sentence utterances.

Other Analyses

Since release of the "standard scoring software" used for the results reported at this meeting, NIST has developed additional scoring software tools. One of these tools performs an analysis of the results reported for each lexical item.

By focussing on individual lexical items ("words") we can investigate lexical coverage as well as performance for individual words for each individual test (such as the June 1990 test). In this RM2 test set there were occurrences of 226 mono-syllabic words and 503 polysyllabic words -- larger coverage of the lexicon than in previous test sets. The most frequently appearing word was "THE", with 297 occurrences.

In the case of the system we refer to as "BBN (2400 train)" with the word pair grammar, in the case of the word "THE" -- 97.6% of the occurrences of this word were correctly recognized, with 0.0% substitution errors, 2.4% deletions, and 0.7% "resultant insertions", for a total of 3.0% word error for this lexical item. What we term "resultant insertions" correspond to cases for which an insertion error of this lexical item occurred, but for which the cause is not known.

The conventional scoring software provides data on a "weighted" frequency-of-occurrence basis. All errors are counted equally, and the more frequently occurring words -- such as the "function" words -- typically contribute more to the overall system performance measures. However, when comparing results from one test set to another it is sometimes desirable to look at measures that are not weighted by frequency of occurrence. Our recently developed scoring software permits us to do this, and, by looking at results for the subset of words that have appeared on all tests to date, some measures of progress over the past several years are provided, without the complications introduced by variable coverage and different frequencies-of-occurrence of lexical items in different tests. Further discussion of this is to appear in an SLS Note in preparation at NIST.

By further partitioning the results of such an analysis into those for mono- and poly-syllabic word subsets, some insights can be gained into the state-of-the art as evidenced by the present tests.

For the speaker-dependent systems trained on 2400 sentence utterances using the word-pair grammar, the unweighted total word error for mono-syllabic word subset is between 1.6% and 2.2% (with the MIT/LL system having a slightly (but not significantly) larger number of "resultant insertions". For the corresponding case of poly-syllabic words, the unweighted total word error is 0.2% for each system.

For the CMU speaker independent system, using the word-pair grammar, the unweighted total word error for mono-syllabic words is 5.6%, and for poly-syllabic words, 1.7%.

By comparing the CMU speaker-independent system results to the best-trained speaker-dependent systems, one can observe that the error rates for mono-syllabic words are typically 3 to 4 times greater than for the speaker-dependent systems, and for poly-syllabic words, approximately 8 times larger. When making similar comparisons, using results for other speaker-independent systems and the best-trained speaker-dependent systems, the mono-syllabic word error rates are typically 4 to 6 times greater, and for poly-syllabic words, 12 times larger.

It is clear from such comparisons that the well-trained speaker-dependent systems have achieved substantially greater success in modelling the poly-syllabic words than the speaker-independent systems.

Comparisons With Other RM Test Sets

Several sites have noted that the four speakers of the RM2 Corpus are significantly different from the speakers of the RM1 corpus. One speaker in particular appears to be a "goat", and there may be two "sheep" -- to varying degrees for both speaker-dependent and speaker-independent systems. An ANOVA test should be implemented to address the significance of this effect.

It has been noted that there appears to be a "within-session effect" -- with later sentence utterances being more difficult to recognize than earlier.

It has been argued that overall performance is worse for this test set than for other recent test sets in the RM corpora, but this conclusion does not appear to be supported for all systems. Some sites have noted that performance for this test set is worse than for the RM2 Development Test Set, but the significance of this effect is unknown. Data for the current AT&T system are available for both the Feb 89 and Oct 89 Speaker Independent Test Sets, and indicate total word errors of 5.2% and 4.7%, respectively (see

Table 2) vs. 5.7% for the June 1990 RM2 test set (see Table 1), suggesting that the RM2 test set is more difficult. A similar comparison involving the current CMU data for the Feb 89 and Oct 89 Speaker Independent Test Sets indicates word error rates of 4.6% and 4.8%, respectively vs. 4.3% for the June 1990 test set, suggesting that for the current CMU system there is (probably insignificantly) better performance on the June 1990 test set. The significance of these differences is not known, but appears to vary from system to system.

Summary

This paper has presented NIST's tabulation and preliminary analysis of results reported for DARPA Resource Management benchmark speech recognition tests just prior to the June 1990 DARPA Speech and Natural Language Workshop at Hidden Valley, PA. The results are provided for both speaker-dependent, speaker-adaptive, and speaker-independent systems, using both RM2 and RM1 test material. All results reported in this document were scored at NIST using NIST scoring software. The reader is referred to other papers in the Proceedings (e.g., references [4 - 9]) for details of the systems and additional discussion of these results.

Acknowledgements

We would like to acknowledge the cooperation of the following individuals who served as points-of-contact (and in many cases, principal researcher) at their site: Jay Wilpon (AT&T), Francis Kubala (BBN), Kai-Fu Lee (CMU), Doug Paul (MIT/LL), Hy Murveit (SRI), and (for SSI), Bill Meisel and Kai-Fu Lee. At NIST, Jon Fiscus has been responsible for development and implementation of scoring tools, and John Garofolo has been responsible for production of the Resource Management Corpora on CD-ROM.

References

[1] "DARPA Extended Resource Management Continuous Speech Speaker-Dependent Corpus (RM2)", NIST speech discs 3-1.1 and 3-2.1, April 1990.

[2] "DARPA Resource Management Continuous Speech Database (RM1)", NIST speech discs 2-1.1/2-2.1, 2-3.1, and 2-4.1, 1989-1990.

[3] Pallett, D. S., "Tools for the Analysis of Benchmark Speech Recognition Tests", paper S2.16 in Proceedings of ICASSP 90, International Conference on Acoustics, Speech and Signal Processing, April 3-6, 1990, pp. 97-100.

[4] Paul, D. B., "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

[5] C. H. Lee et al., "Improved Acoustic Modeling for Continuous speech Recognition", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

[6] Kubala, F. and Schwartz, R., "A New Paradigm for Speaker-Independent Training and Speaker Adaptation", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

[7] Huang, X. et al., "Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

[8] Murveit, H. Weintraub, M. and Cohen, M., "Training Set Issues in SRI's DECIPHER Speech Recognition System", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

[9] Anikst, M. T. et al., "Experiments with Tree-Structured MMI Encoders on the RM Task", Proceedings of DARPA Speech and Natural Language Workshop, June 1990.

June 1990 RM2 (Four Speaker) Test Set

Speaker-Dependent Systems

a. Word-Pair Grammar:

	Corr	Sub	Del	Ins	Total Err	Sent Err
BBN (2400 train)	98.5	1.1	0.5	0.1	1.7	11.3
BBN (600 train)	97.3	2.1	0.6	0.4	3.1	20.0
MIT/LL (2400 train)	98.7	0.9	0.4	0.2	1.5	11.0
MIT/LL (600 train)	97.4	1.7	0.9	0.5	3.1	20.0

b. No Grammar:

	Corr	Sub	Del	Ins	Total Err	Sent Err
MIT/LL (2400 train)	95.9	3.3	0.9	0.8	4.9	28.8
MIT/LL (600 train)	89.5	8.3	2.2	2.2	12.7	58.3

Speaker-Independent Systems

a. Word-Pair Grammar, 109-Speaker Training:

	Corr	Sub	Del	Ins	Total Err	Sent Err
AT&T (first run)	94.8	3.9	1.2	0.8	6.0	32.3
AT&T (2nd run/debugged)	94.9	3.7	1.4	0.6	5.7	31.5
CMU	96.2	2.9	0.9	0.5	4.3	27.1
MIT/LL	94.8	3.8	1.3	0.7	5.9	31.9
SRI	94.1	4.8	1.1	0.6	6.5	32.1
SRI (109 + 12 train)	95.6	3.4	0.9	0.4	4.8	27.1
SSI (VQ FE, CI HMM BE)	81.8	11.5	6.7	1.2	19.5	69.8
SSI (SSI FE, CI HMM BE)	85.8	10.4	3.9	1.3	15.6	59.6
SSI (SSI FE, CD HMM BE)	92.4	5.3	2.4	0.4	8.0	41.3

b. No Grammar, 109-Speaker Training:

	Corr	Sub	Del	Ins	Total Err	Sent Err
AT&T (first run)	77.7	16.7	5.6	1.5	23.8	78.3
AT&T (2nd run/debugged)	77.7	16.7	5.6	1.5	23.8	78.3
CMU	81.9	14.8	3.4	1.8	19.9	74.4
MIT/LL	79.1	16.5	4.4	2.1	22.9	74.6
SRI	75.7	18.3	6.0	1.5	25.7	77.3

Table 1.

Results Reported to NIST for Previous Test Sets

a. AT&T (109-speaker training - 2nd run/debugged retest):

	Corr	Sub	Del	Ins	Total Err	Sent Err
AT&T (Feb '89 SI WPG)	95.5	3.4	1.1	0.7	5.2	28.0
AT&T (Feb '89 SI NG)	80.5	15.0	4.5	2.3	21.7	75.3
AT&T (Oct '89 SI WPG)	96.2	2.9	0.9	0.9	4.7	27.3
AT&T (Oct '89 SI NG)	80.6	14.5	5.0	2.6	22.0	76.7

b. BBN (Feb '89 SI set - not previously reported upon):

	Corr	Sub	Del	Ins	Total Err	Sent Err
BBN (Feb '89 SI-12 WPG)	93.7	4.9	1.4	1.1	7.4	37.0
BBN (Feb '89 SI-109* WPG)	94.8	4.3	1.0	1.2	6.5	34.3

(109* => 4360 sentence utterances used for training)

c. BBN (Feb '89 SD set, speaker-adaptive):

	Corr	Sub	Del	Ins	Total Err	Sent Err
BBN (Feb '89 SA-1 WPG)	95.6	3.4	1.0	0.7	5.2	25.7
BBN (Feb '89 SA-4 WPG)	96.4	2.5	1.1	0.7	4.3	23.3

d. CMU (109-speaker training retest):

	Corr	Sub	Del	Ins	Total Err	Sent Err
CMU (Feb '89 SI WPG)	96.1	3.2	0.6	0.7	4.6	24.0
CMU (Oct '89 SI WPG)	96.2	2.7	1.0	1.0	4.8	28.0

e. SSI: (June '88 set, 109 speaker training)

	Corr	Sub	Del	Ins	Total Err	Sent Err
VQ FE, CI HMM BE, WPG	80.3	14.9	4.8	2.2	22.0	71.3
SSI FE, CI HMM BE, WPG	86.3	10.8	2.9	1.5	15.2	55.7
SSI FE, CD HMM BE, WPG	93.6	5.1	1.3	0.7	7.1	36.7

Table 2.

Speaker-Dependent Word-Pair Grammar
Sentence-Level McNemar Test Analysis

	bbn	ll	bbnl	lll
bbn		same 399 27 28 26	bbn 373 53 11 43	bbn 366 60 18 36
ll			ll 358 69 26 27	ll 368 59 16 37
bbnl				same 339 45 45 51
lll				

Legend

bbn => BBN, 2400 training utterances
ll => LL, 2400 training utterances
bbnl => BBN, 600 training utterances
lll => LL, 600 training utterances

Table 3.

Speaker-Independent Word-Pair Grammar Sentence-Level McNemar Test Analysis

	att	att1	cmu	ll	sri	ssi1	ssi2	ssi3
att		same 320 5 9 146	cmu 274 51 76 79	same 268 57 59 96	same 271 54 55 100	att 122 203 23 132	att 163 162 31 124	att 234 91 48 107
att1			same 275 54 75 76	same 268 61 59 92	same 272 57 54 97	att1 126 203 19 132	att1 164 165 30 121	att1 237 92 45 106
cmu				same 275 75 52 78	cmu 272 78 54 76	cmu 134 216 11 119	cmu 173 177 21 109	cmu 248 102 34 96
ll					same 264 63 62 91	ll 129 198 16 137	ll 163 164 31 122	ll 236 91 46 107
sri						sri 129 197 16 138	sri 166 160 28 126	sri 228 98 54 100
ssi1							ssi2 115 30 79 256	ssi3 127 18 155 180
ssi2								ssi3 183 11 99 187
ssi3								

Legend

att => AT&T (first run)
att1 => AT&T (2nd run/debugged)
cmu => CMU
ll => MIT/LL
sri => SRI
ssi1 => SSI (VQ FE - CI HMM BE)
ssi2 => SSI (SSI FE - CI HMM BE)
ssi3 => SSI (SSI FE - CD HMM BE)

Table 4.