1993 BENCHMARK TESTS FOR THE ARPA SPOKEN LANGUAGE PROGRAM

David S. Pallett, Jonathan G. Fiscus, William M. Fisher, John S. Garofolo, Bruce A. Lund, and Mark A. Przybocki

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

ABSTRACT

This paper reports results obtained in benchmark tests conducted within the ARPA Spoken Language program in November and December of 1993. In addition to ARPA contractors, participants included a number of "volunteers", including foreign participants from Canada, France, Germany, and the United Kingdom. The body of the paper is limited to an outline of the structure of the tests and presents highlights and discussion of selected results. Detailed tabulations of reported "official" results, and additional explanatory text appears in the Appendix.

1. INTRODUCTION

Benchmark tests were implemented within the ARPA Human Language Technology research program during the period November 1993 - January 1994. As in tests conducted last year, the large-vocabulary continuous speech recognition technology tests made use of Wall Street Journal-based Continuous Speech Recognition (WSJ-CSR) corpus material which was collected at SRI International (SRI) under contract to the Linguistic Data Consortium (LDC). Spoken language understanding technology tests made use of ARPA Air Travel Information System (ATIS) material collected at several sites, processed at NIST, annotated at SRI, and provided to participating members of the LDC.

2. WSJ-CSR TESTS

2.1. New Conditions

All sites participating in the WSJ-CSR tests were required to submit results for (at least) one of two "Hub" tests. The Hub tests were intended to measure basic speaker-independent performance on either a 64K-word (Hub 1) or 5K-word (Hub 2) read-speech test set, and included required use of either a "standard" 20K trigram (Hub 1) or 5K bigram (Hub 2) grammar, and also required use of standard training sets. These requirements were intended to facilitate meaningful cross-site comparisons.

The "Spoke" tests were intended to support a number of different challenges.

Spokes 1, 3 and 4 supported problems in various types of adaptation: incremental supervised language model adaptation (Spoke 1), rapid enrollment speaker adaptation for "recognition outliers" (i.e., non-native speakers) (Spoke 3), incremental speaker adaptation (Spoke 4). [There were no participants in what had been planned as Spoke 2.]

Spokes 5 through 8 supported problems in noise and channel compensation: unsupervised channel compensation (Spoke 5), "known microphone" adaptation for two different microphones (Spoke 6), unsupervised channel compensation for 2 different environments (Spoke 7), and use of a noise compensation algorithm with a known alternate microphone for data collected in environments when there is competing "calibrated" noise (radio talk shows or music) (Spoke 8). Spoke 9 included spontaneous "dictation-style" speech.

Additional details are found in Kubala, et al. [1], on behalf of members of the ARPA Continuous speech recognition Corpus Coordinating Committee (CCCC).

2.2. WSJ-CSR Summary Highlights

The design of the "Hub and Spoke" test paradigm, was such that opportunities abounded for informative contrasts (e.g., the use of bigram vs. trigram grammars, the enablement/disablement of supervised vs. unsupervised adaptation strategies, etc).

There were nine participating sites in the Hub 1 tests and five sites participating in the Hub 2 tests, and some sites reported results for more than one system or research team.

The lowest word error rate in the Hub 1 baseline condition was achieved by the French CNRS-LIMSI group [2,3]. Application of statistical significance tests indicated that the performance differences between this system and a system developed by Cambridge University Engineering Department using the "HMM Toolkit" approach [4-6], were not significant. The Cambridge University HMM Toolkit approach also yielded excellent results for the smaller-vocabulary Hub 2 tests. The lowest word error rate for an ARPA contractor on the Hub 1 test data, for the C1 condition permitting valid cross-site comparisons, was reported by the group at CMU [7-9]. The CMU results were not significantly different from the corresponding results for the Cambridge University HMM Toolkit system. The lowest word error rate for an ARPA contractor for the (less constrained) P0 condition was reported by the group at BBN.

It is difficult to summarize results of the spoke tests, except to note that there were results reported for 8 different "spoke conditions", with from 1 to 3 participants and systems typically involved in each spoke. Details are presented in the Appendix.

2.3. WSJ-CSR Discussion

In NIST's analyses of the results, displays of the range of reported word error rates for each speaker across all systems are sometimes informative. These displays tend to draw attention to particularly problematic speakers or systems. Figure 1 shows data for the 10 speakers and 11 systems participating in the required Hub 1 C1 test. The speakers have been ordered from low error rate at the top of the figure to high error rate at the bottom. The length of the plotted line indicates the range in word error rate reported over all systems, and the one-standard-deviation points about the mean are indicated with a "+" symbol.

Note that three speakers (40h, 40j, and 40f) have unusually high error rates relative to the other seven in this test set.

In previous tests involving the Resource Management Corpus, it was noted that high error rates seemed to be correlated, at least indirectly, with unusually fast or slow rate of speech. To see if this was the case for the present test data, NIST obtained estimates of the average speaking rate (words/minute) for each of the test speakers. These estimates were based solely on the total number of words uttered and the total duration of the waveform files, and more sophisticated measures would be desirable. Figure 2 shows a plot of the word error rate vs. speaking rate for the 10 speakers and 11 systems in the Hub 1 C1 test.

This figure, like Figure 1, indicates that speakers 40h, 40j and 40f not only have unusually high error rates relative to the other speakers in this test set, but it also indicates that for these speakers, the speaking rate is markedly higher than for the other seven. Whereas the speaking rate for the seven speakers ranges from approximately 115 to 145 words/minute, for the three speakers with high error rate, the speaking rate ranges from 165-175 words/minute.

There are at least two factors that may contribute to higher error rates at these fast speaking rates: within-word and across-word coarticulatory effects (e.g., phone deletions) associated with fast (possibly better described as "careless" or "casual") speech, and possible under-representation of these effects in the training material.

Chase, et al. [9], at CMU, noted that for the 4 speakers in Spoke 7 (40g, 40h, 40i, and 40j), two (40g and 40i) could be subjectively characterized as "careful speaker[s]", but that 40h was characterized as a "pretty fast speaker, [with] very low gain", and 40j as a "very, very fast speaker". These "fast speakers" appear in a number of the test sets.

NIST's analyses of the distributions of rate of speech for two sets of training material for the Hub 1 test (each consisting of approximately 30,000 utterances: "short-term" and "long-term" speakers) indicate that the distributions are rather broad, with the short-term speakers' distribution peaking at 130 words/minute, with a standard deviation of 30 words/minute, and the long-term speakers' distribution peaking at 145 words/minute, with an associated standard deviation of 30 words/minute. Note that speaking rates for the 3 "fast-talking" speakers fall just outside the "plus one standard deviation region" range relative to the peak of the distribution for the "short-term speaker" training set, and just inside the corresponding region relative to the "long-term" training set.

Because a number of the measured performance differences between systems were small, and the results of the paired-comparison significance tests validated the relevant null hypotheses, it has been observed that, in general, the use of larger test sets, especially for the Hub tests, would have been more informative, especially with regard to the results of significance tests requiring larger speaker populations (i.e., the Sign and Wilcoxon Signed-Rank tests). With larger populations of test speakers, it would be less likely to have such disproportionately large representation of "fast speakers" in the test sets.

Two spokes made use of microphones other than the "standard" Sennheiser close-talking microphone. (See, for example, the discussion in the Appendix of this paper for Spokes 5 and 6.) Too other spokes dealt with the issue of performance degradations that were presumably due to degradations in the signal-to-noise ratio. (See, for example, the discussion for Spokes 7 and 8.)

For the test data of Spokes 5-7, subsequent to the completion of the tests, NIST performed signal-to-noise ratio (SNR) analyses, using three different bandwidth (signal pre-processing) conditions: broadband, A-weighted, and 300 Hz-3000 kHz passband "telephone bandwidth". The filtered SNR's are generally higher than the broadband values. Figure 3 shows the results of these SNR analyses.

Figure 3 (a) indicates the SNRs measured for the data of Spoke 5, which includes 10 "unknown" microphones in addition to the simultaneously collected reference Sennheiser close-talking microphone data for each data subset, collected in the normal data collection environment. SRI's "normal offices for recording" speech data have A-weighted sound level values in the 46-48 dB range. There were 2 "tieclip" or lapel microphones, 5 stand-mounted microphones, a surface-effect microphone, a speakerphone, and a cordless telephone in this set of 10 test microphones.

Note that the SNR values for the Sennheiser microphone are typically about 45 dB for the both the broadband and A-weighted conditions, indicating that there is little low-frequency energy in the spectrum of the noise in the Sennheiser microphone data. Sennheiser microphone data typically yield values of 50 dB for the telephone-bandwidth condition. For the alternate microphones, the broadband SNR's range from about 23 dB (for the Audio-Technica stand-mounted microphone) to 45 dB (for the GE cordless telephone). With filtration the SNR's are higher, as expected. Note that nearly all of the microphones provide at least a 30 dB telephone-bandwidth SNR, and that the AT Pro 7a lapel-mounted microphone provides approximately 40 dB.

Figure 3 (b) indicates the measured SNR's for the data of Spoke 6, which includes 2 "known" alternate microphones in addition to the reference Sennheiser close-talking microphone, collected in the normal data collection environment. For the Sennheiser close-talking microphone, the broadband SNR's are, as for Spoke 5, 45-46 dB. There is a substantial difference between the broadband and A-weighted SNRs for the Audio-Technica stand-mounted microphone, corresponding to low frequency noise picked up by this microphone, and for the telephone-bandwidth condition the SNR is approximately 35 dB. With the telephone handset, SNRs are 38 to 40 dB, depending on bandwidth.

The test set data for Spoke 7, shown in Figure 3 (c), involved use of two different microphones (an Audio-Technica stand-mounted microphone and a telephone handset in addition to the usual "reference" Sennheiser close-talking microphone), in two different noise environments, with background A-weighted noise levels of 58-68 dB.

In the quieter of the two "noisy" environments, a computer laboratory with a reported A-weighted sound level in the 58-59 dB range, the broadband SNR was approximately 34-36 dB for the Sennheiser microphone, and 35 dB for the telephone handset data, but only 17 dB for the Audio-Technica microphone. Spectral analyses of the Audio-Technica background noise data demonstrate the presence of significant low frequency energy as well as the presence of harmonic components with an approximately 70 Hz fundamental. These components may have originated in some rotating machinery (e.g., a cooling fan or disc drive).

In the noisier environment, a room containing machinery with conveyor belts for sorting packages, with a reported A-weighted sound level in the 62-68 dB range, the broadband SNR ratio for the Sennheiser data degraded to 27-29 dB (a decrease of approximately 7 dB), and 27 dB for the telephone handset data, and the Audio-Technica to 16 dB (a decrease of only 1 dB). With A-weighting, in the quieter environment, the SNR for the Sennheiser improved very slightly (less than 1 dB, relative to the broad band values), and for the Audio-Technica it was 25 dB, 8 dB higher than the broad band value.

In the noisier environment, the A-weighted S/N ratio for the Sennheiser data was approximately 29 dB, and the Audio-Technica 20 dB.

For the telephone handset data, both the telephone-bandwidth-filtered and the A-weighted SNRs were higher than, but typically within one or two dBs, of the unweighted values, as might be expected.

In summary, for the quieter of the two environments used in collecting the data of Spoke 7, none of the data subsets in Spoke 7 had an average filtered SNR worse than about 25 dB, and in the noisier environment, the worst average filtered SNR for any data subset was approximately 20 dB. These SNR values would not ordinarily be regarded as indicative of severe noise-degradation.

Spoke 8 involved data collected in the presence of competing noise -- music and talk radio broadcasts. For the case of competing music, the broadband SNR for the reference Sennheiser microphone ranged from 44 DB for the so-called "20 dB" condition, to 36 dB for the "10 dB" condition, and 29 dB for the "0 dB" condition. For the Audio-Technica microphone, corresponding measured values were 25, 17, and 11 dB. NIST's measurements of SNR for the data containing competing speech were inconclusive because of the difficulty of distinguishing between the spoken test material and the competing talk radio.

3. ATIS TESTS

3.1. New Conditions

Recent ATIS tests were similar in many respects to previous ATIS tests -- the primary difference consisting of expansion of the size of the relational air-travel-information database to 46 cities, and use of a body of newly collected and annotated data using this relational database [10]. As in prior years, tests included spontaneous speech recognition (SPREC) tests, natural language understanding (NL) tests and spoken language understanding (SLS) tests. For the first time, data collected at NIST was included in the test and training data. The NIST data was collected using systems provided to NIST by BBN and SRI.

In previous years, results for NL and SLS tests were presented and discussed in terms of a "weighted error" percentage, which was computed as twice the percentage of incorrect answers plus the percentage of "No Answer" responses. The decision to weight "wrong answers" twice as heavily as "no answer" responses was reconsidered within the past year by the ARPA Program Manager, and this year only unweighted NL and SLS errors are reported (i.e., incorrect answers count the same as "No Answer" responses). For most system developers, this change of policy has appeared to result in changed strategies for system responses, so that in this year's reported results, little use was made of the "No Answer" response.

3.2. Summary Highlights

For the recent ATIS tests, results were reported for systems at seven sites. Lowest error rates were reported by the group at CMU [11]. The magnitude of the differences between systems is frequently small, and the significance of these small differences is not known.

As in previous years, error rates for "volunteers" are generally higher than for ARPA contractors, possibly reflecting a lesser level-of-effort.

Additional details about the test paradigm, and comments on some aspects by individual participants, are found in another paper in this Proceedings, by Dahl, et al., on behalf of members of the ARPA Multi-site ATIS Data COllection Working (MADCOW) Group [10]. Details about the technical approaches used by the participants, and their own analyses and comments, are to be found in references [11,23-28].

3.3. ATIS Discussion

This year, 46% of the utterances were classified as Class A and 34% in Class D, so that 80% of the test utterances were "answerable" (i.e., Class A or D). Last year's test set had about the same percentage of Class A queries (43%), but somewhat fewer classified as Class D (i.e., 25%), so that last year only 67% were answerable. One possible reason for this change (other than the test-set-to-test-set fluctuations) may be that the Principles of Interpretation document is continually being extended to cover phenomena that would have otherwise resulted in categorization of some queries as "unanswerable", and therefore Class X.

For text input (NL test), for last year's test material, the lowest unweighted NL error rate was 6.5% for the Class A+D subset, 6.5% for Class A, and 6.4% for Class D, in contrast with this year's corresponding figures of 9.3%, 6.0% and 13.8%. Note that this year's test set apparently had "more difficult" Class D queries, and that there was a larger fraction of the queries that were classified as Class D than last year (34% vs. 25%).

For speech input (SLS test), and for last year's unweighted test material, the unweighted SLS error rate was 11.0% for the Class A+D subset, 10.2% for Class A, and 12.5% for Class D, in contrast with this year's corresponding figures of 13.2%, 8.9% and 17.5%.

Note that while the lowest error rate for Class A queries is smaller this year (i.e., 8.9% vs. 10.2%), this year's best Class D error rate was substantially higher than last year's. It may be the case that this is related to the extended coverage provided by the current Principles of Interpretation document, so that queries that in previous years would have been classified as unanswerable, are now judged to be answerable, although context-dependent.

4. ACKNOWLEDGEMENTS

The "Hub and Spokes" Test paradigm could not have been developed, specified, or implemented without the tireless and effective efforts of Francis Kubala, as Chair of the ARPA continuous speech recognition Corpus Collection Coordinating Committee (CCCC). The tests would also not have been possible without the dedicated efforts of Denise Danielson and her colleagues at SRI in collecting an exceptionally large and varied amount of CSR data for CSR system training and test purposes. In the ATIS community, Debbie Dahl served as Chair of the MADCOW group, and it is to her credit that new data was collected at several sites with the 46 city relational database and that participating sites reached agreement on the details of the current tests. Kate Hunicke-Smith and her colleagues at SRI International were again responsible for annotation of ATIS data and for assisting NIST in the adjudication process following preliminary scoring. It is a pleasure to acknowledge Kate's thoughtful and cheerful interactions with our group at NIST.

As in previous years, the cooperation of many participants in the ARPA data and test infrastructure -- typically several individuals at each site -- is gratefully acknowledged.

REFERENCES

[1] Kubala, F., et al., "The Hub and Spoke Paradigm for CSR Evaluation", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.).

[2] Gauvain, J.L., Lamel, L.F., Adda, G. and Adda-Decker, M., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task", in Proceedings of ICASSP'94.

[3] Gauvain. J.L., Lamel. L.F. Adda, G. and Adda-Decker, M., "The LIMSI Continuous Speech Dictation System" in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.).

[4] Woodland, P.C., Odell, J.J., Valtchev, V. and Young, S.J., "Large Vocabulary Continuous Speech Recognition Using HTK", in Proceedings of ICASSP'94.

[5] Odell, J.J., Woodland, P.C., and Young, S.J., "Tree-based State Tying for High Accuracy Acoustic Modelling", in Proceedings of the Human Language Technology Workshop, March 1994, (Weinstein, C.J., ed.).

[6] Odell, J.J., Valtchev, V., Woodland, P.C., and Young, S.J., "A One Pass Decoder Design for Large Vocabulary Recognition," in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.).

[7] Hwang, M., Thayer, E. and Huang, X., "Semi-Continuous HMMs with Phone-Dependent VQ Codebooks for Continuous Speech Recognition" in Proceedings of ICASSP'94.

[8] Hwang, M., et al., "Improving Speech Recognition Performance via Phone-Dependent VQ codebooks and Adaptive Language Models in SPHINX-II" in Proceedings of ICASSP'94.

[9] Hwang, M., Thayer, E., Mosur, R. and Chase, L., "Phone-Dependent Codebooks and Multiple Speaker Clusters in SPHINX-II", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[10] Dahl, D., et al., "Expanding the Scope of the ATIS Task: The ATIS-3 Corpus", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.).

[11] (a) Ward, W. and Issar, S., "Recent Improvements in the CMU Spoken Language Understanding System", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.), and (b) Issar, s., and ward, W., "Flexible Parsing: CMU's Approach to Spoken Language Understanding", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[12] Garofolo, J., Robinson, T. and Fiscus, J., "The Development of File Formats for Very Large Speech Corpora: SPHERE and Shorten", in Proceedings of ICASSP'94.

[13] (a) Zavaliagkos, G., et al., "BBN Hub System and Results", (b) Lapre, C., et al., "Speaker Adaptation for Non-Native Speakers", (c) Anastasakos, A. et al., "Environmental Robustness: Adaptation to Known Alternate Microphones", and (d) Nguyen, L. et al., "Spoke 9: Spontaneous WSJ Dictation", Oral Presentations at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[14] Ostendorf, M. et al., "Stochastic segment Modelling for Continuous Speech Recognition: Wall Street Journal Benchmark Report", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[15] (a) Scattone, F., et al., "Dragon's Large Vocabulary Speech Recognition System", (b) Orloff, J., et al., "Spoke S4: Speaker Adaptation", and (c) Orloff, J., et al., "Spoke S6: Microphone Adaptation", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[16] Morgan, N., et al., "Scaling a Hybrid HMM/MLP System for Large Vocabulary CSR", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[17] (a) Paul, D.B., "The Lincoln Large Vocabulary Stack-Decoder Based HMM CSR", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.), and (b) Paul, D.B., "The Lincoln Large Vocabulary Stack-Decoder Based HMM CSR: Spoke S4 Incremental Speaker Adaptation", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[18] (a) Robinson, T., Hochberg, M. and Renals, S., "IPA: Improved Phone Modelling with Recurrent Neural Networks", in Proceedings of ICASSP'94, (b) Hochberg, M., Robinson, T., and Renals, S. "ABBOT: The CUED Hybrid Connectionist-HMM Large-Vocabulary Recognition System", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[19] Aubert, X., et al., "The Philips Large Vocabulary CSR System", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[20] Aubert, X., Dugast, C., Ney, H. and Steinbiss, V., "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", in Proceedings of ICASSP'94.

[21] (a) Rosenfeld, R., "A Hybrid Approach to Adaptive Statistical Language Modelling", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.), and (b) Chase, L., Mosur, R., and Rosenfeld, R., "Language Model Adaptation in the CSR Evaluation", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[22] (a) Liu, F.H., Moreno, P.J., Stern, R.M., and Acero, A., "Signal Processing for Robust Speech Recognition", in Proceedings of the Human Language Technology Workshop, March 1994 (Weinstein, C.J., ed.) and (b) Stern, R.M., Liu, F.H., and Moreno, P., "Robust Speech Recognition: Research at CMU", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[23] Bocchieri, E., "The ATT ATIS System: March 94 Report", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[24] (a) Stallard, D., et al., "Recent Work in Spoken Language Understanding in the BBN SLS Project", and (b) Miller, S., et al., "Statistical Language Processing Using Hidden Understanding Models", Oral Presentations at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[25] Normandin, Y., "CRIM's December 1983 ATIS System", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[26] "The MIT ATIS System: March 1994 Progress Report", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[27] Moore, R. and Cohen, M. et al., "SRI's Recent Progress on the ATIS Task", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[28] Dahl, D., Linebarger, M., Nguyen, N. and Norton, L., "Unisys Activities in Spoken Language Understanding", Oral Presentation at the Spoken Language Technology Workshop, March 6-8,

[29] Digilakis, V., et al., "SRI November 1993 CSR Hub evaluation", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

[30] Weintraub, M., et. al., "SRI November 1993 CSR Spoke Evaluation", Oral Presentation at the Spoken Language Technology Workshop, March 6-8, 1994, Princeton, NJ.

NOTICE

Throughout this paper, a number of references are provided in order to refer readers to relevant papers and oral presentations by researchers at the individual sites participating in the tests. In some of these papers, results are cited that differ by small amounts from those tabulated in this paper. In some cases the authors cite unofficial or preliminary, "pre-adjudication" results. In other cases, the authors cite other unofficial test results conducted after the "official" test period closed.

The views expressed in this paper are those of the author(s). The results presented are for local, system-developer-implemented tests. NIST's role in the tests is one of selecting and distributing the test materials, implementing scoring software, and uniformly tabulating the results of the tests. The views of the author(s) and these results are not to be construed or represented as endorsements of any systems or official findings on the part of NIST, ARPA or the U.S. Government.

APPENDIX: "BENCHMARK TEST RESULTS"

A.1. WSJ-CSR November 1993 Test Material

The 1993 WSJ-CSR tests make use of newly-collected training material, a new compressed waveform file format, new test paradigms, and new test sets.

The new training material for the WSJ-CSR task includes a substantial amount of data (31 CD-ROMs containing training and developmental test data) collected at SRI International under contract to the Linguistic Data Consortium (LDC).

In a collaborative effort involving NIST, Tony Robinson at Cambridge University's Engineering Department, and the LDC, the newly collected waveform data was processed with an "embedded" version (i.e., the file's SPHERE-format header is uncompressed, but the bulk of the file is compressed) of a lossless waveform compression algorithm ("shorten") using the NIST SPHERE file header convention, to reduce the storage requirements for this data by a factor of approximately 50% [12]. The CSR test material was released in November.

A.2. WSJ-CSR Test Scoring and Adjudication

The CSR tests were conducted in November and December. Test and scoring protocols were similar to last year. However, new to the CSR benchmark tests this year was the addition of an official adjudication period. Following a preliminary scoring of recognition results, sites participating in the tests were permitted to submit requests for adjudication to NIST. Adjudication requests in the CSR domain contained requests for transcription modifications due to transcription errors, alternative transcriptions, etc.

A total of 22 bug reports were received from 6 sites. The bug reports contained requests for changes to 199 (151 unique) utterance transcriptions in all WSJ-CSR test sets. The NIST adjudicators carefully evaluated each request and ultimately revised transcriptions of 83 utterances (55% of the ones in question.)

Of the transcriptions that were revised, most were the result of judgements by the adjudicators that the transcriptions contained words which could have multiple orthographic representations (e.g., compound words, variant orthographic representations, etc.) or which were lexically ambiguous. In many of these cases, both the original transcription and an alternative transcription were permitted. This was implemented by mapping alternate word forms to a single form in both the transcriptions and the recognized strings. The remaining revisions were the correction of simple transcription errors.

A.3. WSJ-CSR Test Participants

United States participants in the WSJ-CSR tests included: BBN Systems and Technologies (BBN) [13], Boston University (BU) [14], Carnegie Mellon University (CMU) [7-9], Dragon Systems [15], the International Computer Science Institute (ICSI) at Berkeley [16], Massachusetts Institute of Technology's Lincoln Laboratory (MIT/LL) [17], and SRI International (SRI) [29,30].

Foreign participants included two British groups at Cambridge University's Engineering Department, one pursuing connectionist approaches (CU-CON) [18], and another, developers of the HMM Toolkit (CU-HTK) [4-6], a French group at CNRS-LIMSI (LIMSI) [2,3], and a German group at the Philips GmbH Research Laboratories in Aachen [20].

BU collaborated with BBN, making use of the N-best outputs of a BBN system, using an N-best rescoring formalism, a stochastic segment modelling approach, and the use of both BU and BBN knowledge sources.

A.4. WSJ-CSR Benchmark Test Results

A.4.1. Hub 1: 64K Baseline. The intention of the two "Hub" tests was "to improve basic [speaker independent] performance on clean [read speech] data". For Hub 1, test data consisted of 200 utterances -- 20 from each of 10 speakers, using the primary (Sennheiser series HMD 410) microphone as used in prior tests.

All sites were required to provide results for a static (i.e., non-adaptive) Speaker-Independent (SI) baseline system that would permit cross-site comparisons, which would use the standard 20K word trigram "open vocabulary" grammar and use standardized training sets.

The results of that baseline system are tabulated in the column labelled "Contrast C1" in Table 1.

Results for (optional) use of the same system training, but with the 20K bigram grammar, are shown in the column labelled "Contrast C2". These 'contrastive' results were intended for comparison with results for optional 'primary' systems. The primary systems could use "any grammar or acoustic training", and these results are shown in the column labelled "P0".

In most cases, data from each site shows on a single line. The three BU "C1" systems each represent different N-best rescoring formalisms using the BU stochastic segment model recognition system in combination with the BBN Byblos system, using different knowledge sources to re-rank the N-best hypotheses. The two different CMU systems are different in many ways, so that comparisons are non-trivial.

For the baseline "C1" systems, word error rates ranged from 19.0% to 11.7%, with the lowest error rate reported for the LIMSI system.

In this table, and others of this sort in this paper, the results of contrastive comparisons are shown in the boxes labelled "COMPARISONS AND SIGNIFICANCE TESTS". The results of use of the NIST statistical significance tests that have been used in previous tests are also shown.

To illustrate interpretation of some of the tabulated results, note that BBN and MIT/LL achieved reductions in error rate of 13.9% and 9.8%, respectively, for their P0 systems when compared to the C1 baseline systems. In most cases, these reductions were shown to be significant. Refer to [13] and [17] for discussion of factors contributing to these reduction error rate.

When contrasting use of trigram and bigram grammars, a number of sites achieved reductions in error rate of from approximately 12% to 23% for the case of use of the trigram grammar.

Table 2 shows a matrix tabulation of the results of cross-site and, in some cases, within-site, paired comparison statistical significance tests for the baseline H1-C1 systems.

A.4.2. Hub 2: 5K Baseline. Because run times for full 20K systems were in some cases regarded as prohibitive, a second baseline Hub test, requiring only a 5K lexicon, was permitted. For Hub 2, the required static SI baseline C1 system made use of a standard 5K bigram closed vocabulary grammar and either of two smaller training sets, consisting of approximately 7200 sentence utterances.

As for Hub 1, the Hub 2 test data consisted of 200 utterances -- 20 from each of 10 speakers, using the primary microphone.

Not surprisingly, error rates for the 5K systems were lower than for the 20K systems.

Table 3 shows that for the baseline C1 systems, error rates ranged from 17.7% to 8.7%, with the lowest error rate reported by the Cambridge University's HTK research group [4-6]. For the P0 systems, for which "any grammar or acoustic training" were permissible, lower error rates were to be expected, and were achieved, typically with reductions in error rate of from 25% to almost 50%. In this case, also, one of the HTK configurations achieved the lowest word error rate: 4.9%.

Table 4 shows a matrix tabulation of the results of cross-site and, in some cases, within-site, paired comparison statistical significance tests for the baseline H2-C1 systems.

A.4.3. Spoke 1: Language Model Adaptation. The stated goal for this language model adaptation spoke was "to evaluate an incremental supervised language model (LM) adaptation algorithm on a problem of sublanguage adaptation". The sole participant was Rosenfeld et al. at CMU [21]. Test data consisted of read speech data from four speakers, each reading 1 to 5 articles consisting of approximately 20-25 sentence utterances, with the Sennheiser microphone. NIST's scoring was done on four successive 5-sentence utterance blocks throughout the articles (i.e., utterances 1-5, 6-10, 11-15, and 16+). Use of the statistical significance tests was not thought to be appropriate since these tests assume independence of errors across sentences, and this assumption is probably not valid when using an adaptive language model.

Table 5 presents the results for Spoke 1. The column labelled P0 shows results with incremental unsupervised adaptation enabled: word error rates vary from 16.5% on the first block of 5 sentences to 18.2% on the last block. In contrast, with language model adaptation disabled, the word error rates correspondingly vary from 20.5% to 21.1%. Comparisons between P0 and C1, involving enabling/disabling of supervised LM adaptation, indicate reductions in word error rate of between 9.8% to 19.4%, with lesser reductions for the P0:C2 comparisons involving unsupervised LM adaptation.

A.4.4. Spoke 3: SI Recognition Outliers. The stated goal for this spoke was "to evaluate a rapid enrollment speaker adaptation algorithm on difficult speakers (e.g., non-native speakers of American English)". The sole participant was BBN [13]. Test data consisted of read speech from ten speakers, each reading 40 sentence utterances, with the Sennheiser microphone. For each speaker, the 40 "rapid enrollment" utterances were available for use with the "rapid enrollment" speaker adaptation.

Table 6 presents the results for Spoke 3. The column labelled P0 shows results with rapid enrollment adaptation enabled: word error rate for the 400 utterance test set is 14.5%. In contrast, with adaptation disabled, the word error rate is 32.0%. Alternatively, the P0:C1 contrast indicates a reduction in error rate 54.7%, which was shown to be significant using all of the significance tests applied by NIST.

A.4.5. Spoke 4: Incremental Speaker Adaptation. The stated goal for this spoke was "to evaluate an incremental speaker adaptation algorithm". Two sites participated: Dragon [15] and MIT/LL [17]. In this spoke, there were only four test speakers, with 100 sentence utterances for each. NIST's scoring was done on four successive 25-sentence utterance blocks (i.e., utterances 1-25, 26-50, 51-75, and 76+).

Table 7 presents the results for Spoke 4.

For the Dragon results, word error rates for the P0 condition (with incremental unsupervised adaptation enabled) range from 15.5% to 14.3%. For MIT/LL, the corresponding variation is 10.9% to 11.1%. There is evidence of significant reductions in error of the order of 20% to 30% for the P0:C1 contrasts for the Dragon results (e.g., note the reduction of from 19.4% to 15.5% for the first block of 25 utterances).

For the corresponding MIT/LL results, the magnitudes of the reductions are not as large. For both sites, the incremental changes in error rates between the P0 and C2 cases, involving unsupervised/supervised adaptation, in most cases are not shown to be significant, and range from approximately 4% to 16%.

A.4.6. Spoke 5: "Microphone Independence". The stated goal of this spoke was to "evaluate an unsupervised channel compensation algorithm". The different "channels" in this case were different microphones -- each of the ten speakers in this test set used a different (unknown) microphone. Similar, but not identical, microphones had been incorporated in training and development material. For the 200 utterances in each portion of this test set, both the unknown microphone data (in "wv2" data files) and corresponding Sennheiser microphone data (in "wv1" files) were available.

Both CMU [22] and SRI [30] participated in this spoke.

Table 8 presents the results for Spoke 5.

With unsupervised channel compensation enabled, the CMU system achieved an error rate of 15.1%, in contrast to 20.9% with compensation disabled -- a 27.8% reduction in word error rate. SRI achieved a comparable reduction of 24.2%, and with slightly lower error rates. With compensation enabled, the CMU system achieved 9.7% word error for the corresponding Sennheiser data, while the SRI system achieved 6.6% word error. Enabling/disabling the channel compensation made essentially no difference for the case of the Sennheiser data subset, as might be suspected.

A.4.7. Spoke 6: Known Alternate Microphones. The stated goal of this spoke was to "evaluate a known microphone adaptation algorithm". There were two different microphones -- an Audio Technica stand-mounted microphone, and a telephone handset which was to be connected to the data collection apparatus "over external lines", in addition to the Sennheiser (wv1) data. Two-channel microphone adaptation data -- for each of the two microphones and the (reference) Sennheiser microphone was provided from "devtest data". There were ten speakers for the data for each of the two microphones, with 20 sentence utterances per speaker. In NIST's analysis of the results, data are separately tabulated for the Audio-Technica (at) data, and for the telephone handsets (th).

Three sites participated: BBN [13], Dragon [15], and SRI [30].

Table 9 presents the results for Spoke 6.

For the case of the microphone adaptation disabled (C1), for the Audio-Technica microphone's data, word error rates were 6.4% for the SRI system, 10.4% for the BBN system, and 18.5% for the Dragon results. For telephone handset data, the SRI system had 19.1%, the BBN system had 29.3% and Dragon 65.4%. These results for the telephone handset data were probably somewhat worse than might have been expected because of inadvertent channel differences between development test and evaluation test sets.

Considering the adaptation enabled/disabled P0:C1 contrast, BBN and Dragon achieved 9.4% and 11.7% reductions in word error rate for the Audio-Technica microphone, and 57.4% (from 29.3% to 12.5% word error) and 11.7% for BBN and Dragon, respectively. On corresponding Sennheiser data, the BBN and SRI systems with adaptation disabled achieved word error rates ranging from 5.9% to 8.4%, while the Dragon results were 13.8% and 14.6%.

A.4.8. Spoke 7: "Noisy Environments". The stated goal of this spoke was to "evaluate a noise compensation algorithm with known alternate microphones" in two different data-collection environments with background A-weighted sound level of from 55 to 68 dB. Two different microphones were used, the same microphones as were used for Spoke 6, (the Audio-Technica and a telephone handset). Utterances for the microphone/channel adaptation (Sennheiser to known alternate microphone) were available from development test data, and there were files with background noise (but no speech) for each microphone-noise-environment-speaker condition. The two noise environments ("e1" and "e2") consisted of computer laboratory (e1), and a room with package sortation machinery in operation ("e2").

The sole participant in this spoke was SRI [30].

Table 10 presents the results for Spoke 7.

As might be expected, the word error rate was smallest for the lower of the two noise conditions with the alternate high-quality (but not close-talking) Audio-Technica microphone (8.5%) (for which the A-weighted S/N ratio was approximately 26 dB), and markedly higher for both alternate microphones in the higher noise environment (17.4% and 28.8%). For corresponding data from the close-talking Sennheiser microphone, in the two different noise environments, error rates of from 6.3% to 9.1% were obtained.

A.4.9. Spoke 8: "Calibrated Noise Sources". The stated goal of this spoke was to "evaluate a noise compensation algorithm with a known alternate microphone on data corrupted with calibrated noise sources". Data was collected using the Audio-Technica microphone, which was also used in Spokes S6 and S7, in the presence of competing noise (from a "boom box" radio-tape player situated nearby). The competing noise was either a variety of musical selections ("mu") or talk radio ("tr"). The competing noise was "calibrated" in the sense that the level of the competing noise was intended to be set so as to be 20 or 10 dB less than the speech peak level, or equal to (or potentially greater than) the speech peak level, the "0 dB condition". Note however that NIST's measurements of SNR do not agree well with these desiderata, as discussed in Section 2.3 of this paper except in some qualitative sense.

CMU [22] was the sole participant in this spoke.

Table 11 presents the results for Spoke 8.

Data were submitted for the 3 competing noise conditions, both microphones (Sennheiser and Audio-Technica), and with noise compensation enabled and disabled -- a total of 24 conditions, permitting many cross-comparisons.

With compensation disabled, there were reductions in error rate with use of the close-talking, noise cancelling Sennheiser microphone when comparing results for the two different microphones (C3:C1). With compensation enabled, and again comparing the two different microphones (C3:P0), the differences in error rate are reduced, but are still significant in most cases.

There is evidence of significant reductions in error rate when considering compensation enabled/disabled (P0:C1) for both music and talk radio at the 10 dB and 0 dB conditions.

Further, enabling compensation appears to be beneficial for much of the data obtained with the close talking Sennheiser microphone (see, for example the C3:C2 comparisons).

A.4.10. Spoke 9: Spontaneous WSJ Dictation. The stated goal of this spoke was to "improve basic performance on spontaneous dictation-style speech". There were 10 speakers (all journalists, but with varying experience in dictation), each dictating 20 spontaneous Wall Street Journal-like sentence utterances, and using the Sennheiser microphone.

BBN [13] was the sole participant in this spoke.

Table 12 presents the results for Spoke 9.

Using the same system as used for the C1 condition in Hub 1 (which achieved a word error rate of 14.2% on the Hub 1 test data), a word error rate of 24.7% was achieved on the S9 data, indicating that the spontaneous dictation S9 test set is substantially more challenging. BBN's S9 system achieved an error rate of 19.1% on the S9 data, a significant reduction in word error rate of 22.8% over the H1-C1 system.

A.5. ATIS November 1993 Test Material

The final, adjudicated set of test material consisted of 965 test utterances and was collected at 5 sites -- BBN, CMU, MIT, NIST and SRI. As in previous years, it was selected by NIST staff from set-aside material previously collected within the MADCOW community [10]. The test set was selected so as to balance the number of utterances per data collection site (~200 utterances per site.) Because of differences in the scenarios and data collection systems used at the different collection sites, it was not possible to balance the test set for number of subjects or the difficulty of scenarios per collection site. No "pre-filtering" of the test data was performed except to attempt to exclude subject-scenarios with mostly repetitive queries. The ATIS test material was released in November, 1993.

A.6. ATIS Scoring and Adjudication

The ATIS scoring and adjudication process took place in December and early January. ATIS test and scoring protocols were similar to those of previous benchmark tests. After the scored ATIS results were released in December 1993, approximately 140 adjudication requests ("bug reports") were sent to NIST. NIST worked in conjunction with SRI to resolve the requests, about 10 of which were duplicates.

The majority of the bug reports dealt with transcription issues, in some cases pointing to limitations in our community's procedures for transcribing ATIS-domain spontaneous speech. One utterance, in particular, which was classified as Class X (and thus did not affect the NL or SLS scores), but was included in the ATIS SPREC scoring, included low-level remarks by the experimenter, as a result of an inadvertent "open mike" condition. Originally, this block of speech was transcribed as "unintelligible", but in adjudication, it was fully transcribed, partially because a number of sites had objected to having been scored with significant numbers of insertion errors. After adjudication, most sites continued to do very poorly on this one utterance, but were now penalized for substitutions and deletions as well. It alone accounts for an increment of approximately 0.3% in the Class A+D+X word error for most sites, and a substantially larger fraction of the Class X error rate. In retrospect, it is clear that this problematic utterance (and the entire subject-scenario) ought not to have been included in the test set because of the "open mike" condition.

Besides the recurrent complaints of bad transcriptions, a problem involving fare IDs or flight IDs not appearing in the maximal reference answer files (the "rf2s") (which came to be known as "Joe's Fare Bug") was brought to our attention. This bug was attributed to about 21 of the test utterances before scoring. The bug was fixed by SRI and new .rf2s were generated prior to rescoring.

A.7. ATIS Test Participants

United States participants in the ATIS tests included: AT&T Bell Laboratories (AT&T) [23], BBN Systems and Technologies (BBN) [24], Carnegie Mellon University (CMU) [11], Massachusetts Institute of Technology's Laboratory for Computer Science (MIT/LCS) [26], and SRI International (SRI) [27], and Unisys (UNISYS) [28]. There was one foreign participant: (CRIM) [25], from Canada.

AT&T collaborated with CMU, using an AT&T-developed

ATIS-domain speech recognition system and the CMU ATIS natural language system, and Unisys collaborated with BBN, using a set of N-best outputs for a BBN ATIS-domain speech recognition system as input for Unisys-developed natural language technology.

A.8. ATIS Benchmark Test Results

A.8.1. SPontaneous speech RECognition (SPREC) Tests. Table 13 presents the results for the SPREC tests for all systems and subsets of the ATIS test data, using the Sennheiser close-talking microphone. For the case of the subset of all answerable queries, Class A+D, the word error rates ranged from 3.3% to 9.0%.

Table 14 presents a matrix tabulation of the ATIS SPREC results for the Class A+D subset. The overall word error rate across all tested systems for the data from the several collecting sites ("Overall Totals" row along the bottom of the Table) ranges from 3.6% for the CMU-collected data to 6.8% for the NIST-collected data, reflecting differences in subject populations and other factors.

Table 15 presents the results, in matrix form, of the application of 4 paired-comparison significance tests for the SPREC systems for the Class A+D subset. Among other things, note that the performance differences between the BBN and the CMU systems are not shown to be significant, and that the differences between the MIT, SRI and one of the Unisys systems are also not shown to be significant. Note also that significant differences are shown between the BBN results and those for the two Unisys systems, which make use of BBN-provided N-best results.

A.8.2. Natural Language (NL) Understanding Tests. Table 16 presents a tabulation of the results for the NL tests for all systems and all sets of "answerable" ATIS queries, Class A+D, Class A and Class D.

For the set of all answerable queries, Class A+D, the unweighted error rate ("UW. Err.") ranges from 43.1% to 9.3%. For Class A queries, the range is 28.6% to 6.0%, and for Class D, the range is 63.1% to 13.8%. In each case (and as in last year's results), the lowest error rates were reported by the CMU system.

As noted in Section A9 of this paper, the AT&T NL system was the results of a collaborative agreement with CMU, thus it is not surprising that the performance is nearly identical to that of the CMU system.

There are, in some cases, more than one set of results submitted by individual sites, corresponding to different systems. The differences between systems were specified in the "Systems Descriptions" provided to NIST at the time results were submitted. Space limitations prohibit discussion of these differences in this paper.

After preliminary scoring had been completed, Moore at SRI advised NIST that a bug had been found in the code that produced results submitted to NIST for the SRI NL and SLS systems, with the effect of reporting results that were "essentially the output of [the SRI] system with the robust processing component turned off", because a "No_Answer" response over-wrote the answer produced by the robust processing component (a "template matcher"). With the permission of the ARPA Coordinating Committee, SRI later resubmitted results for the debugged systems, and these SRI results are shown as "late, debugged" results.

Table 17 presents a matrix tabulation of the official NL results for the several subsets of test material. There is some indication of varying degrees of difficulty presented by the different subsets of data from the different sites, subject-scenarios, and subject populations: note that the unweighted error rates reported in the "Overall Totals" row ranges from 28.1% to 16.0%, but also note that both these values were obtained with BBN systems -- one at BBN, and the other at NIST. These differences probably are not significant since the numbers of speakers in the individual test sets is small.

A.8.3. Spoken Language System (SLS) Understanding Tests. Table 18 presents a tabulation of the results for the SLS tests for all systems and all sets of "answerable" ATIS queries, Class A+D, Class A and Class D.

For the set of all answerable queries, Class A+D, the unweighted error rate ("UW. Err.") ranges from 46.8% to 13.2%. For Class A queries, the range is 33.5% to 8.9%, and for Class D, the range is 65.2% to 17.5%. For the Class A+D and Class A results, the lowest error rates were obtained by the CMU system, but for the Class D results, the lowest error rates were obtained by the MIT/LCS system.

Table 19 presents a matrix tabulation of the official SLS results for the several subsets of Class A+D test material from different sites. Note that there is some evidence of "local adaptation" to locally collected data (e.g., error rates for the CMU system are substantially lower for the CMU-collected data).

Note also that some sites (typically the "volunteers") continued to use the "No_Answer" option more frequently than others, which would be a beneficial strategy in a system in which "wrong answers" were penalized more heavily than "no answer". In some cases, use of this option was more prevalent for data from some originating sites than others, perhaps reflecting differences between subject populations or subject-scenario subsets.

RANGE ANALYSIS ACROSS SPEAKERS FOR THE TEST:
November 1993 Hub 1, Contrast 1
by Speaker Word Error for Speakers

```
|-----------------------------------------------------------------------------------------------|
|        |                                  PERCENTAGES                                         |
|--------+--------------------------------------------------------------------------------------|
|        |0    5    10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95   100|
|  SPKR  ||    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    ||
|--------+--------------------------------------------------------------------------------------|
| 40A    |     +-|-+--                                                                          |
| 40I    |     --+-|-+---                                                                       |
| 40G    |        +|-+                                                                          |
| 40C    |       +--|--+---                                                                     |
| 40E    |         +-|-+---                                                                     |
| 40B    |          +|-+-                                                                       |
| 40D    |          +--|--+                                                                     |
| 40H    |                +--|---+---                                                           |
| 40J    |                       -+--|---+-----                                                 |
| 40F    |                     --+----|---+--                                                   |
|-----------------------------------------------------------------------------------------------|

        | -> shows the mean
        + -> shows plus or minus one standard deviation
```

Figure 1 Range of Word error rates for the 10 speakers of the Hub 1 C1 test set for 11 systems

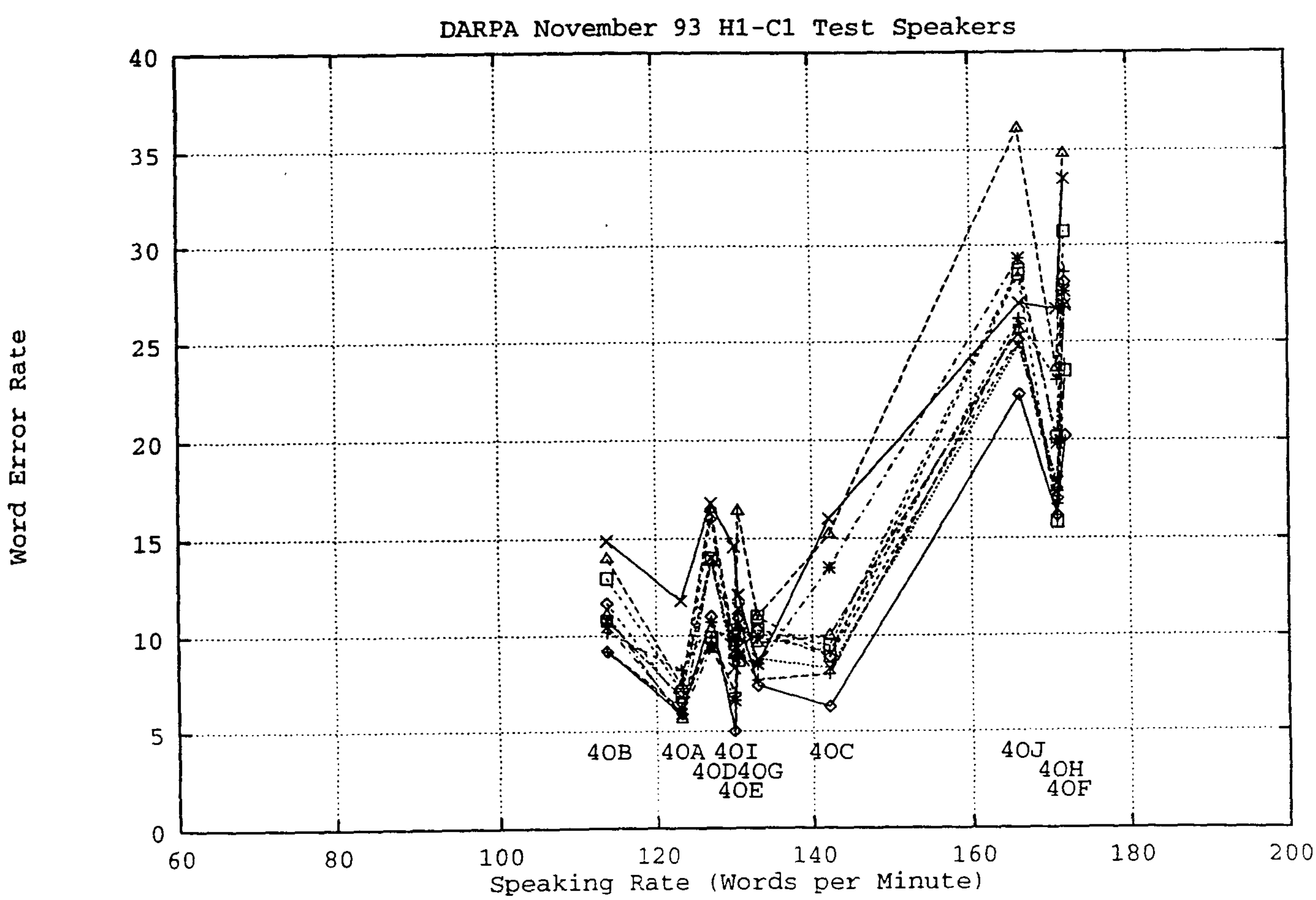


Figure 2 Word error rates for Hub 1 C1 speakers vs. speaking rate

Fig. 3A: SNR Measurements for Spoke 5

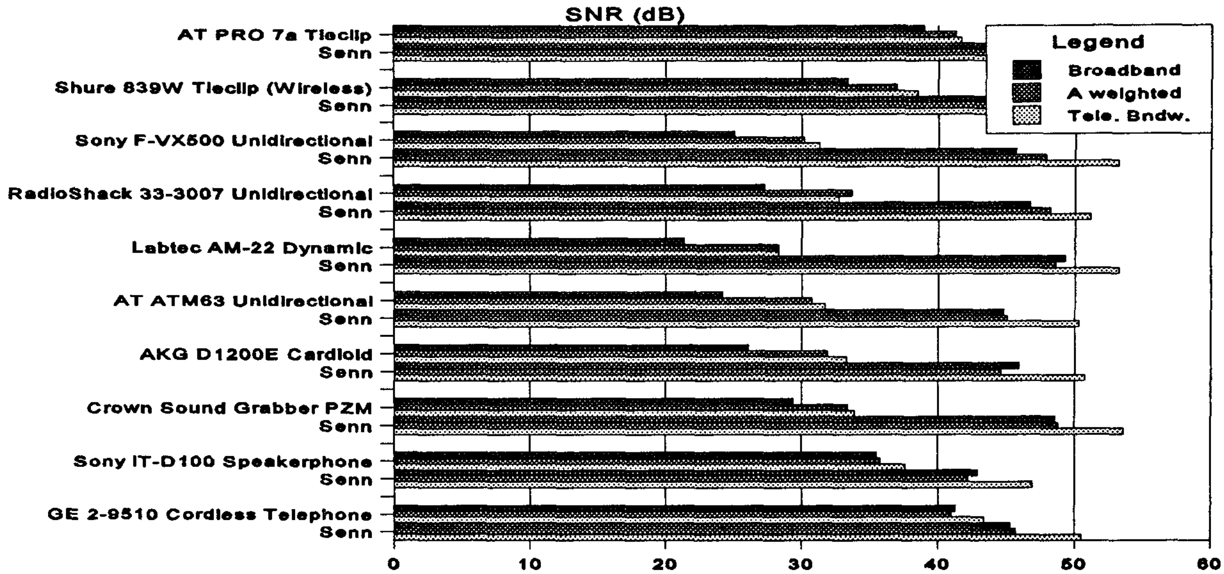


Fig. 3B: SNR Measurements for Spoke 6

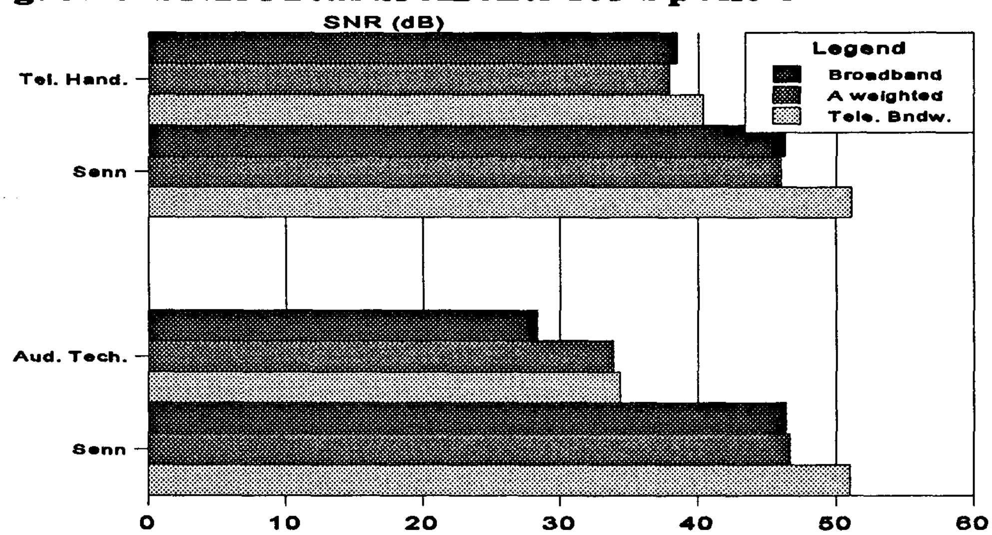
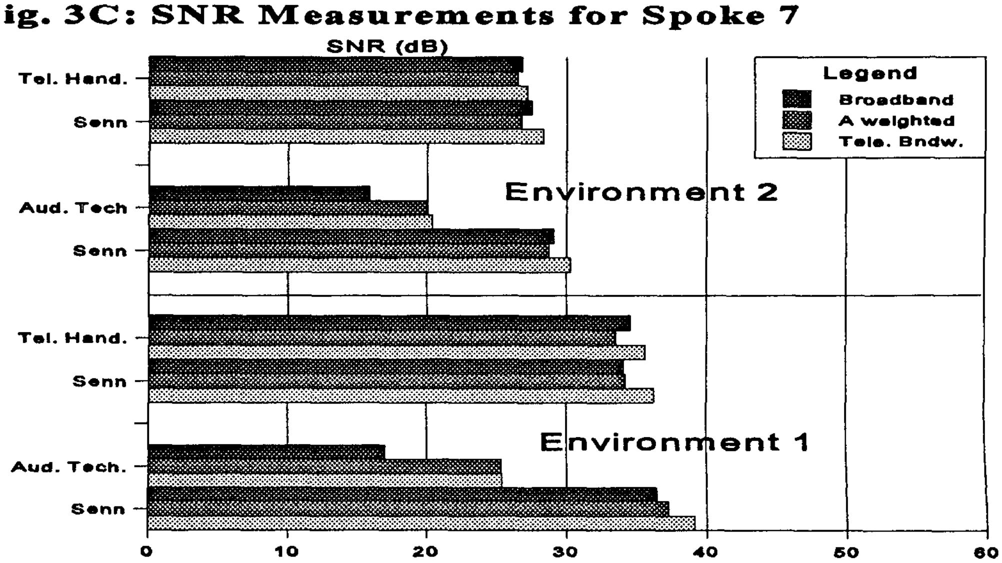


Fig. 3C: SNR Measurements for Spoke 7



Nov 93 Hub and Spoke CSR Evaluation
Hub 1: 64K Read WSJ Baseline

GOAL: improve basic SI performance on clean data.
DATA: 10 speakers * 20 utts = 200 utts 64K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (opt) any grammar or acoustic training, session boundaries and utterance order given as side information.

C1 (req) Static SI test with standard 20K trigram open-vocab grammar and choice of either short-term or long-term speakers

C2 (opt) Static SI test with standard 20K bigram open-vocab grammar and choice of either short-term or long-term speakers

SIDE INFO: Session boundaries and utterance order are known for H1-P0 only.

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn1	12.2	14.2	
bu1		15.7	
bu2		14.3	
bu3		14.5	
cmu1		13.6	
cmu2	13.9		
cu-htk1		12.7	14.4
dragon1		19.0	
limsi1		11.7	15.2
mit-ll1	16.8	18.6	
philips2		14.8	17.2
sri1		14.4	16.5

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn1	P0:C1	13.9%	P0	same	P0	P0
mit-ll1	P0:C1	9.8%	P0	P0	P0	P0

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
cu-htk1	C1:C2	11.7%	C1	C1	C1	same
limsi1	C1:C2	22.7%	C1	C1	C1	C1
philips2	C1:C2	14.0%	C1	C1	C1	same
sri1	C1:C2	13.0%	C1	C1	C1	C1

Table 1 Hub 1 Results

Composite Report of All Significance Tests
For the h1_c1 Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy) Test	WI
McNemar (Sentence Error) Test	MN

	bbn1-h1_c1	bu1-h1_c1	bu2-h1_c1	bu3-h1_c1	cmu1-h1_c1	cu-htk1-h1_c1	dragon1-h1_c1	limsi1-h1_c1	mit-111-h1_c1	philips2-h1_c1	sri1-h1_c1
bbn1-h1_c1		MP bbn1-h1_c1 SI same WI bbn1-h1_c1 MN bbn1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same
bu1-h1_c1			MP bu2-h1_c1 SI same WI bu2-h1_c1 MN same	MP bu3-h1_c1 SI same WI bu3-h1_c1 MN same	MP cmu1-h1_c1 SI same WI cmu1-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN same	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP bu1-h1_c1 SI same WI bu1-h1_c1 MN same	MP same SI same WI same MN same	MP same SI same WI same MN same
bu2-h1_c1				MP same SI same WI same MN same	MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI same WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN same	MP limsi1-h1_c1 SI same WI limsi1-h1_c1 MN limsi1-h1_c1	MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN same	MP same SI same WI same MN same	MP same SI same WI same MN sri1-h1_c1
bu3-h1_c1					MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP bu3-h1_c1 SI bu3-h1_c1 WI bu3-h1_c1 MN same	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP bu3-h1_c1 SI bu3-h1_c1 WI bu3-h1_c1 MN bu3-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same
cmu1-h1_c1						MP same SI same WI same MN same	MP cmu1-h1_c1 SI same WI cmu1-h1_c1 MN cmu1-h1_c1	MP limsi1-h1_c1 SI same WI same MN limsi1-h1_c1	MP cmu1-h1_c1 SI same WI cmu1-h1_c1 MN cmu1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same
cu-htk1-h1_c1							MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI same WI same MN same
dragon1-h1_c1								MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP same SI same WI same MN same	MP philips2-h1_c1 SI philips2-h1_c1 WI philips2-h1_c1 MN philips2-h1_c1	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
limsi1-h1_c1									MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP limsi1-h1_c1 SI same WI limsi1-h1_c1 MN limsi1-h1_c1	MP limsi1-h1_c1 SI same WI same MN limsi1-h1_c1
mit-111-h1_c1										MP philips2-h1_c1 SI philips2-h1_c1 WI philips2-h1_c1 MN philips2-h1_c1	MP sri1-h1_c1 SI same WI sri1-h1_c1 MN sri1-h1_c1
philips2-h1_c1											MP same SI same WI same MN same
sri1-h1_c1											

Table 2 Significance Test Results: Hub 1 (64K Baseline) C1 Systems

Nov 93 Hub and Spoke CSR Evaluation
Hub 2: 5K Read WSJ Baseline

GOAL: improve basic SI performance on clean data.
DATA: 10 speakers * 20 utts = 200 utts 5K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (opt) any grammar or acoustic training, session boundaries and utterance order given as side information.

C1 (req) Static SI test with standard 5K bigram closed-vocab grammar and choice of either short-term or long-term speakers from WSJ0 (7.2K utts).

SIDE INFO: session boundaries and utterance order are known for H2-P0 only.

	Primary P0	Contrast C1
System	Word Err. (%)	Word Err. (%)
bu1	6.7	11.6
bu2	5.4	10.3
bu3	5.8	10.8
cu-con1		13.5
cu-htk2	4.9	8.7
cu-htk3		12.5
icsi1		17.7
limsi2	5.2	9.3
philips1	9.2	12.3
philips2	6.4	

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bu1	P0:C1	42.4%	P0	P0	P0	P0
bu2	P0:C1	47.4%	P0	P0	P0	P0
bu3	P0:C1	46.6%	P0	P0	P0	P0
cu-htk2	P0:C1	43.4%	P0	P0	P0	P0
limsi2	P0:C1	43.7%	P0	P0	P0	P0
philips1	P0:C1	25.5%	P0	P0	P0	P0

Table 3 Hub 2 Results

Composite Report of All Significance Tests
For the h2_c1 Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy) Test	WI
McNemar (Sentence Error) Test	MN

	bu1-h2_c1	bu2-h2_c1	bu3-h2_c1	cu-con1-h2_c1	cu-htk2-h2_c1	cu-htk3-h2_c1	icsi1-h2_c1	limsi2-h2_c1	philips1-h2_c1
bu1-h2_c1		MP bu2-h2_c1 SI same WI bu2-h2_c1 MN same	MP same SI same WI same MN same	MP bu1-h2_c1 SI same WI same MN same	MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1	MP same SI same WI same MN same	MP bu1-h2_c1 SI bu1-h2_c1 WI bu1-h2_c1 MN bu1-h2_c1	MP limsi2-h2_c1 SI limsi2-h2_c1 WI limsi2-h2_c1 MN limsi2-h2_c1	MP same SI same WI same MN same
bu2-h2_c1			MP bu2-h2_c1 SI same WI bu2-h2_c1 MN same	MP bu2-h2_c1 SI same WI same MN same	MP cu-htk2-h2_c1 SI same WI same MN cu-htk2-h2_c1	MP bu2-h2_c1 SI same WI same MN same	MP bu2-h2_c1 SI bu2-h2_c1 WI bu2-h2_c1 MN bu2-h2_c1	MP same SI limsi2-h2_c1 WI same MN limsi2-h2_c1	MP bu2-h2_c1 SI same WI same MN same
bu3-h2_c1				MP bu3-h2_c1 SI same WI same MN same	MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1	MP bu3-h2_c1 SI same WI same MN same	MP bu3-h2_c1 SI bu3-h2_c1 WI bu3-h2_c1 MN bu3-h2_c1	MP limsi2-h2_c1 SI limsi2-h2_c1 WI same MN limsi2-h2_c1	MP bu3-h2_c1 SI same WI same MN same
cu-con1-h2_c1					MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1	MP same SI same WI same MN same	MP cu-con1-h2_c1 SI cu-con1-h2_c1 WI cu-con1-h2_c1 MN cu-con1-h2_c1	MP limsi2-h2_c1 SI limsi2-h2_c1 WI limsi2-h2_c1 MN limsi2-h2_c1	MP same SI philips1-h2_c1 WI philips1-h2_c1 MN same
cu-htk2-h2_c1						MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1	MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1	MP same SI same WI same MN same	MP cu-htk2-h2_c1 SI cu-htk2-h2_c1 WI cu-htk2-h2_c1 MN cu-htk2-h2_c1
cu-htk3-h2_c1							MP cu-htk3-h2_c1 SI cu-htk3-h2_c1 WI cu-htk3-h2_c1 MN cu-htk3-h2_c1	MP limsi2-h2_c1 SI limsi2-h2_c1 WI limsi2-h2_c1 MN limsi2-h2_c1	MP same SI same WI same MN same
icsi1-h2_c1								MP limsi2-h2_c1 SI limsi2-h2_c1 WI limsi2-h2_c1 MN limsi2-h2_c1	MP philips1-h2_c1 SI philips1-h2_c1 WI philips1-h2_c1 MN philips1-h2_c1
limsi2-h2_c1									MP limsi2-h2_c1 SI same WI limsi2-h2_c1 MN limsi2-h2_c1
philips1-h2_c1									

Table 4 Significance Test Results: Hub 2 (5K Baseline) C1 Systems

Nov 93 Hub and Spoke CSR Evaluation
Spoke 1: Language Model Adaptation

GOAL: evaluate an incremental supervised LM adaptation algorithm on a problem of sublanguage adaptation.

DATA: 4 A spkrs * 1-5 articles (~100 utts) = 400 utts. Read unfiltered WSJ data from 1990 publications in TIPSTER corpus, Sennheiser mic, minimum of 20 sentences per article.

Primary and Contrast Conditions

P0 (req) incremental supervised LM adaptation, closed vocabulary, any LM trained from 1987-89 WSJ0 texts

C1 (req) S1-P0 system with LM adaptation disabled

C2 (opt) incremental unsupervised LM adaptation

SIDE INFO: session boundaries and utterance order are known

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
cmu3 Utts 1-5	16.5	20.5	16.9
cmu3 Utts 6-10	17.3	19.2	18.3
cmu3 Utts 11-15	15.9	18.4	16.7
cmu3 Utts 16+	18.2	21.1	18.9

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.
cmu3 Utts 1-5	P0:C1	19.4%
cmu3 Utts 6-10	P0:C1	9.8%
cmu3 Utts 11-15	P0:C1	13.6%
cmu3 Utts 16+	P0:C1	14.0%

	Test Comp.	% Change W.E.
cmu3 Utts 1-5	P0:C2	2.1%
cmu3 Utts 6-10	P0:C2	5.3%
cmu3 Utts 11-15	P0:C2	5.0%
cmu3 Utts 16+	P0:C2	3.9%

	Test Comp.	% Change W.E.
cmu3 Utts 1-5	C2:C1	17.7%
cmu3 Utts 6-10	C2:C1	4.8%
cmu3 Utts 11-15	C2:C1	9.1%
cmu3 Utts 16+	C2:C1	10.5%

Table 5 Spoke 1: Language Model Adaptation Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 3: SI Recognition Outliers

GOAL: evaluate a rapid enrollment speaker adaptation algorithm on difficult speakers.

DATA: 10 B spkrs * 40 utts = 400 utts (test)

10 B spkrs * 40 utts = 400 utts (rapid enrollment from S3 speakers, used for S3-P0)

10 A spkrs * 40 utts = 400 utts (rapid enrollment from Hub speakers, used for S3-C2)

5K-word read WSJ data, Sennheiser mic, collected from non-native speakers of American English (British, European, Asian dialects, etc.).

Primary and Contrast Conditions

P0 (req) rapid enrollment speaker adaptation

C1 (req) S3-P0 system with speaker adaptation disabled

C2 (req) S3-P0 system on H2 data

SIDE INFO: speaker identity is known for P0, C1, and C2, session boundaries and utterance order is known for C3.

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn2	14.5	32.0	10.7

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn2	P0:C1	54.7%	P0	P0	P0	P0

Table 6 Spoke 3: SI Recognition Outlier Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 4: Incremental Speaker Adaptation

GOAL: evaluate an incremental speaker adaptation algorithm.

DATA: 4 A spkrs * 100 utts = 400 utts (test)

A spkrs * 40 utts = 160 utts (rapid enrollment from A speakers in S3)

5K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (req) incremental unsupervised speaker adaptation

C1 (req) S4-P0 system with speaker adaptation disabled

C2 (opt) incremental supervised adaptation

SIDE INFO: for all conditions: session boundaries and utterance order are known; additional for C2: correct transcription is known after the fact.

System	Primary P0 Word Err. (%)	Contrast C1 Word Err. (%)	Contrast C2 Word Err. (%)
dragon2 Utts 1-25	15.5	19.4	14.5
dragon2 Utts 26-50	15.3	19.6	14.5
dragon2 Utts 51-75	15.4	20.3	13.3
dragon2 Utts 76+	14.3	20.9	12.0
mit-112 Utts 1-25	10.9	11.6	11.4
mit-112 Utts 26-50	10.7	11.6	10.2
mit-112 Utts 51-75	10.9	10.4	9.6
mit-112 Utts 76+	11.1	12.0	11.0

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	McN
dragon2 Utts 1-25	P0:C1	20.3%	P0	same
dragon2 Utts 26-50	P0:C1	21.7%	P0	same
dragon2 Utts 51-75	P0:C1	24.3%	P0	same
dragon2 Utts 76+	P0:C1	31.5%	P0	same
mit-112 Utts 1-25	P0:C1	5.8%	same	same
mit-112 Utts 26-50	P0:C1	7.4%	same	same
mit-112 Utts 51-75	P0:C1	-4.7%	same	same
mit-112 Utts 76+	P0:C1	7.0%	same	same

	Runtime Ratio s4_p0/s4_c1
dragon2	0.914
mit-112	1.026

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	McN
dragon2 Utts 1-25	C2:P0	6.2%	same	same
dragon2 Utts 26-50	C2:P0	5.2%	same	same
dragon2 Utts 51-75	C2:P0	13.9%	C2	same
dragon2 Utts 76+	C2:P0	16.1%	C2	C2
mit-112 Utts 1-25	C2:P0	-4.1%	same	same
mit-112 Utts 26-50	C2:P0	4.8%	same	same
mit-112 Utts 51-75	C2:P0	11.8%	C2	same
mit-112 Utts 76+	C2:P0	1.4%	same	same

Table 7 Spoke 4: Incremental Speaker Adaptation Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 5: Microphone Independence

GOAL: evaluate an unsupervised channel compensation algorithm

DATA: 10 A spkrs * 20 utts = 200 utts (2 channels, same speech as H2

5K-word read WSJ data), 10 different mics not in training or development test. NOTE: No speech from the test microphones can be used.

Primary and Contrast Conditions

P0 (req) unsupervised channel compensation enabled on wv2 data

C1 (req) S5-P0 system with compensation disabled on wv2 data

C2 (req) S5-P0 system on Sennheiser (wv1) data

C3 (opt) S5-C1 system on Sennheiser (wv1) data

SIDE INFO: Microphone identities are not known

System	Primary P0 Word Err. (%)	Contrast C1 Word Err. (%)	Contrast C2 Word Err. (%)	Contrast C3 Word Err. (%)
cmu4	15.1	20.9	9.7	9.7
sri2	13.1	17.2	6.6	6.6

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
cmu4	P0:C1	27.8%	P0	same	P0	same
sri2	P0:C1	24.2%	P0	same	P0	same
cmu4	C2:P0	36.1%	C2	C2	C2	C2
sri2	C2:P0	49.5%	C2	C2	C2	C2
cmu4	C2:C1	53.8%	C2	same	C2	C2
sri2	C2:C1	61.7%	C2	C2	C2	C2
cmu4	C3:P0	35.9%	C3	C3	C3	C3
sri2	C3:P0	49.5%	C3	C3	C3	C3
cmu4	C3:C1	53.7%	C3	same	C3	C3
sri2	C3:C1	61.7%	C3	C3	C3	C3
cmu4	C2:C3	0.3%	same	same	same	same
sri2	C2:C3	0.0%	same	same	same	same

Table 8 Spoke 5: "Microphone Independence" Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 6: Known Alternate Microphone

GOAL: evaluate a known microphone adaptation algorithm.

DATA: 10 A spkrs * 20 utts * 2 mics = 400 utts (test, 2 channels)
10 D spkrs * 40 utts * 2 mics = 800 utts (mic-adaptation from devtest, 2channels)

5K-word read WSJ data, from an Audio-Technica directional stand-mounted mic and telephone handset over external lines, plus stereo mic adaptation data. NOTE: the 800 stereo microphone adaptation utterances will come from the devtest and are the only data from the target mics that are the only data from the target mics that is allowed. The only data available for adaptation to the environment will be from the S7 Spoke of the devtest data.

Primary and Contrast Conditions

P0 (req) supervised mic adaptation enabled on wv2 data

C1 (req) S6-P0 system with mic adaptation disabled on wv2 data

C2 (req) S6-C1 system on Sennheiser (wv1) data

SIDE INFO: Microphone identities are known. Use of the stereo mic-adaptation data will be allowed for the S6-P0 condition only

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn3 at	9.4	10.4	7.7
bbn3 th	12.5	29.3	8.4
dragon3 at	16.4	18.5	13.8
dragon3 th	25.3	65.4	14.6
sri2 at		6.4	5.9
sri2 th		19.1	7.2

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn3 at	P0:C1	9.9%	P0	same	P0	same
bbn3 th	P0:C1	57.4%	P0	P0	P0	P0
dragon3 at	P0:C1	11.7%	P0	same	same	P0
dragon3 th	P0:C1	61.3%	P0	P0	P0	P0

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn3 at	C2:P0	18.3%	C2	C2	C2	C2
bbn3 th	C2:P0	32.7%	C2	C2	C2	C2
dragon3 at	C2:P0	15.5%	C2	same	same	C2
dragon3 th	C2:P0	42.3%	C2	C2	C2	C2

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn3 at	C2:C1	26.4%	C2	C2	C2	C2
bbn3 th	C2:C1	71.3%	C2	C2	C2	C2
dragon3 at	C2:C1	25.4%	C2	C2	C2	C2
dragon3 th	C2:C1	77.7%	C2	C2	C2	C2
sri2 at	C2:C1	7.7%	same	same	same	same
sri2 th	C2:C1	62.5%	C2	C2	C2	C2

Table 9 Spoke 6: Known Alternate Microphones Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 7: Noisy Environments

GOAL: evaluate a noise compensation algorithm with known alternate mic.

DATA: 10 A spkrs * 10 utts * 2 mics * 2 envs = 400 utts (test, 2 channels)

5K-word read WSJ data, same 2 secondary mics as in S6, collected in two environments with a background A-weighted noise level of about 55-68 Db. NOTE: the 800 stereo microphone adaptation utterances will come from the devtest and are the only data from the target mics that is allowed. The only data available for adaptation to the environment will be from the S7 Spoke of the devtest data.

Primary and Contrast Conditions

C1 (req) S7-P0 system with compensation disabled on wv2 data

C2 (req) S7-P0 system on Sennheiser (wv1) data

SIDE INFO: Microphone identities are known. Use of the stereo environment-adaptation data will be allowed for the S7-P0 condition only

	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)
sri2 at_e1	8.5	6.3
sri2 at_e2	17.4	9.1
sri2 th_e1	29.1	8.4
sri2 th_e2	28.8	8.3

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
sri2 at_e1	C2:C1	25.0%	C2	same	same	same
sri2 at_e2	C2:C1	47.8%	C2	C2	C2	C2
sri2 th_e1	C2:C1	71.2%	C2	C2	C2	C2
sri2 th_e2	C2:C1	71.4%	C2	C2	C2	C2

Table 10 Spoke 7: "Noisy Environments" Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 8: Calibrated Noise Sources

GOAL: evaluate a noise compensation algorithm with known alternate mic on data corrupted with calibrated *noise sources*

DATA: 10 A spkrs * 10 utts * 2 sources * 3 levels = 600 utts (test, 2 channels)

5K-word read WSJ data collected with competing recorded music or talk radio in the background at 0, 10, and 20 Db SNR, using the Audio-Technica directional stand-mounted mic from S6. *NOTE:* the 400 *stereo microphone adaptation utterances* will come from the devtest and are the only data from the target mic that is allowed.

Primary and Contrast Conditions

P0 (req) noise compensation enabled on wv2 data

C1 (*req*) S8-P0 *system with* compensation disabled on wv2 data

C2 (req) S8-P0 system on Sennheiser (wv1) data

C3 (opt) S8-C1 system on Sennheiser (wv1) data

	Primary P0	Contrast C1	Contrast C2	Contrast C3
System	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)
cmu5 mu_20	14.6	14.8	13.5	11.3
cmu5 mu_10	19.5	32.1	13.4	11.9
cmu5 mu_0	58.7	77.9	18.1	16.9
cmu5 tr_20	15.5	16.3	13.3	12.1
cmu5 tr_10	25.5	36.9	13.4	12.0
cmu5 tr_0	75.2	86.3	25.6	25.6

Nov 93 Hub and Spoke CSR Evaluation
Spoke 8: Calibrated Noise Sources
COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
cmu5 mu_20	P0:C1	1.2%	same	same	same	same
cmu5 mu_10	P0:C1	39.5%	P0	P0	P0	P0
cmu5 mu_0	P0:C1	24.7%	P0	P0	P0	same
cmu5 tr_20	P0:C1	4.9%	same	same	same	same
cmu5 tr_10	P0:C1	30.9%	P0	P0	P0	same
cmu5 tr_0	P0:C1	12.9%	P0	P0	P0	same
	Test Comp.	**% Change W.E.**	**Significance Tests: MAPSSWE**	**Sign**	**Wilcoxon**	**McN**
cmu5 mu_20	C2:P0	7.8%	same	same	same	same
cmu5 mu_10	C2:P0	31.0%	C2	same	C2	same
cmu5 mu_0	C2:P0	69.1%	C2	C2	C2	C2
cmu5 tr_20	C2:P0	14.1%	same	same	same	same
cmu5 tr_10	C2:P0	47.4%	C2	C2	C2	C2
cmu5 tr_0	C2:P0	66.0%	C2	C2	C2	C2
	Test Comp.	**% Change W.E.**	**Significance Tests: MAPSSWE**	**Sign**	**Wilcoxon**	**McN**
cmu5 mu_20	C2:C1	8.8%	same	same	same	same
cmu5 mu_10	C2:C1	58.2%	C2	C2	C2	C2
cmu5 mu_0	C2:C1	76.7%	C2	C2	C2	C2
cmu5 tr_20	C2:C1	18.4%	C2	same	same	same
cmu5 tr_10	C2:C1	63.7%	C2	C2	C2	C2
cmu5 tr_0	C2:C1	70.4%	C2	C2	C2	C2
	Test Comp.	**% Change W.E.**	**Significance Tests: MAPSSWE**	**Sign**	**Wilcoxon**	**McN**
cmu5 mu_20	C3:P0	22.6%	C3	C3	C3	C3
cmu5 mu_10	C3:P0	38.6%	C3	C3	C3	C3
cmu5 mu_0	C3:P0	71.2%	C3	C3	C3	C3
cmu5 tr_20	C3:P0	21.9%	C3	same	same	same
cmu5 tr_10	C3:P0	52.8%	C3	C3	C3	C3
cmu5 tr_0	C3:P0	66.0%	C3	C3	C3	C3
	Test Comp.	**% Change W.E.**	**Significance Tests: MAPSSWE**	**Sign**	**Wilcoxon**	**McN**
cmu5 mu_20	C3:C1	23.5%	C3	same	C3	same
cmu5 mu_10	C3:C1	62.9%	C3	C3	C3	C3
cmu5 mu_0	C3:C1	78.3%	C3	C3	C3	C3
cmu5 tr_20	C3:C1	25.8%	C3	C3	C3	same
cmu5 tr_10	C3:C1	67.4%	C3	C3	C3	C3
cmu5 tr_0	C3:C1	70.4%	C3	C3	C3	C3
	Test Comp.	**% Change W.E.**	**Significance Tests: MAPSSWE**	**Sign**	**Wilcoxon**	**McN**
cmu5 mu_20	C3:C2	16.0%	C3	C3	C3	C3
cmu5 mu_10	C3:C2	11.1%	C3	same	C3	same
cmu5 mu_0	C3:C2	6.8%	same	C3	C3	same
cmu5 tr_20	C3:C2	9.1%	C3	same	C3	C3
cmu5 tr_10	C3:C2	10.2%	C3	C3	C3	same
cmu5 tr_0	C3:C2	0.0%	same	same	same	same

Table 11 Spoke 8: "Calibrated Noise Sources" Results

Nov 93 Hub and Spoke CSR Evaluation
Spoke 9: Spontaneous WSJ Dictation

GOAL: improve basic SI performance on spontaneous dictation-style speech.

DATA: 10 C speakers * 20 utts = 200 utts
Spontaneous WSJ-like dictations (business news stories), Sennheiser mic

Primary and Contrast Conditions

P0 (req) any grammar or acoustic training

C1 (req) S9-P0 system on H1 data

C2 (req) H1-C1 system on S9 data

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn4	19.1	13.7	24.7

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn4	P0:C2	22.8%	P0	P0	P0	P0

Table 12 Spoke 9: Spontaneous WSJ Dictation Results

Dec93 ATIS SPREC Test Results

Class A+D+X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-adx	10.2	91.5	6.2	2.3	1.7	46.3	964
bbn3-adx	4.1	97.0	2.4	0.7	1.0	20.9	964
cmu2-adx	4.1	96.8	2.4	0.9	0.9	22.1	964
crim3-adx	8.3	94.5	4.6	0.9	2.8	36.8	964
mit_lcs2-adx	5.6	95.2	3.3	1.5	0.8	28.4	964
sri3-adx	5.4	95.5	3.0	1.5	0.9	27.5	964
sri4-adx	5.2	95.7	2.8	1.4	0.9	26.0	964
unisys2-adx	5.2	96.1	3.0	0.9	1.3	26.3	964
unisys3-adx	4.9	96.3	2.9	0.9	1.2	23.5	964

Class A+D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a_d	9.0	92.5	5.4	2.1	1.5	42.2	773
bbn3-a_d	3.3	97.5	2.0	0.5	0.8	18.0	773
cmu2-a_d	3.3	97.3	2.0	0.7	0.6	19.7	773
crim3-a_d	6.6	95.7	3.6	0.7	2.3	31.3	773
mit_lcs2-a_d	4.5	96.1	2.6	1.2	0.7	23.3	773
sri3-a_d	4.6	96.1	2.5	1.4	0.7	23.3	773
sri4-a_d	4.5	96.3	2.3	1.3	0.8	22.3	773
unisys2-a_d	4.0	97.0	2.3	0.7	1.0	21.9	773
unisys3-a_d	3.9	97.1	2.3	0.6	0.9	19.7	773

Class A Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a	8.6	93.1	5.0	1.9	1.7	44.4	448
bbn3-a	3.0	97.9	1.7	0.4	0.9	18.5	448
cmu2-a	3.0	97.5	1.8	0.7	0.6	19.9	448
crim3-a	6.3	96.1	3.2	0.7	2.4	31.9	448
mit_lcs2-a	4.3	96.4	2.4	1.1	0.8	24.1	448
sri3-a	4.0	96.6	2.0	1.3	0.7	22.1	448
sri4-a	3.9	96.8	1.9	1.3	0.7	21.0	448
unisys2-a	3.6	97.4	1.9	0.7	1.0	20.8	448
unisys3-a	3.5	97.4	1.9	0.6	1.0	19.6	448

Class D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-d	9.6	91.4	6.1	2.4	1.0	39.1	325
bbn3-d	4.0	96.8	2.6	0.6	0.8	17.2	325
cmu2-d	3.9	96.8	2.5	0.7	0.7	19.4	325
crim3-d	7.2	94.9	4.3	0.7	2.1	30.5	325
mit_lcs2-d	4.9	95.6	3.1	1.4	0.5	22.2	325
sri3-d	5.7	95.2	3.4	1.5	0.9	24.9	325
sri4-d	5.5	95.4	3.2	1.4	0.9	24.0	325
unisys2-d	4.9	96.2	3.2	0.6	1.1	23.4	325
unisys3-d	4.4	96.4	2.9	0.6	0.9	19.7	325

Class X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-x	15.5	87.3	9.7	3.0	2.8	62.8	191
bbn3-x	7.2	94.7	3.9	1.4	1.9	32.5	191
cmu2-x	7.3	94.7	3.8	1.5	2.0	31.9	191
crim3-x	15.0	89.7	8.7	1.6	4.7	59.2	191
mit_lcs2-x	10.0	91.5	5.9	2.5	1.6	49.2	191
sri3-x	8.7	92.8	5.2	1.9	1.6	44.5	191
sri4-x	8.0	93.3	4.9	1.8	1.3	41.4	191
unisys2-x	10.1	92.3	5.9	1.8	2.4	44.5	191
unisys3-x	9.3	93.0	5.2	1.8	2.4	39.3	191

Table 13 ATIS SPREC Benchmark Test Results

Dec93 ATIS SPREC Test Results

Class A+D Subset — Originating Site of Test Data

SYSTEMS	BBN (146 Utt.)	CMU (163 Utt.)	MIT (132 Utt.)	NIST-BBN (89 Utt.)	NIST-SRI (77 Utt.)	SRI (166 Utt.)	Overall Totals 773	Foreign Coll. Site Totals
att2	6.3 1.8 1.3 9.4 49.3	3.4 1.4 1.0 5.8 25.2	4.6 2.3 1.2 8.1 47.0	4.3 3.0 2.9 10.2 51.7	8.6 2.6 2.0 13.2 49.4	6.5 2.0 1.1 9.6 40.4	5.4 2.1 1.5 9.0 42.2	5.4 2.1 1.5 9.0 42.2
bbn3	0.7 0.1 0.5 1.3 7.5	1.4 0.4 0.6 2.4 11.7	2.1 0.3 1.0 3.5 23.5	3.2 1.2 1.9 6.3 34.8	3.6 0.8 0.5 4.9 23.4	2.2 0.4 0.5 3.1 17.5	2.0 0.5 0.8 3.3 18.0	2.3 0.6 0.9 3.8 20.4
cmu2	2.5 0.6 0.4 3.5 23.3	1.3 0.5 0.7 2.5 11.7	2.0 0.5 0.6 3.1 24.2	2.3 1.6 0.9 4.7 24.7	2.3 1.0 0.5 3.8 22.1	2.0 0.4 0.7 3.1 16.9	2.0 0.7 0.6 3.3 19.7	2.2 0.7 0.6 3.5 21.8
crim3	3.0 0.4 1.3 4.8 24.0	2.1 0.7 2.1 4.9 23.3	3.7 0.5 2.5 6.7 34.8	4.4 1.5 3.7 9.6 46.1	5.2 0.8 2.4 8.5 36.4	4.1 0.8 2.3 7.2 32.5	3.6 0.7 2.3 6.6 31.3	3.6 0.7 2.3 6.6 31.3
mit_lcs2	2.2 0.8 0.4 3.4 23.3	1.7 1.5 0.5 3.7 16.6	3.2 1.4 1.0 5.6 31.8	1.9 1.2 0.6 3.9 24.7	4.4 1.1 0.5 6.0 28.6	3.0 1.2 0.7 5.0 19.9	2.6 1.2 0.7 4.5 23.3	2.5 1.2 0.6 4.2 21.5
sri3	2.8 1.0 0.7 4.6 27.4	1.8 1.0 0.6 3.5 14.1	2.6 1.4 0.7 4.6 30.3	2.3 2.0 0.5 4.8 29.2	4.4 1.8 2.0 8.2 31.2	1.9 1.3 0.5 3.8 16.3	2.5 1.4 0.7 4.6 23.3	2.6 1.4 0.8 4.8 25.2
sri4	2.2 1.0 0.5 3.8 23.3	1.8 1.0 0.6 3.4 13.5	2.4 1.2 0.8 4.4 28.0	2.0 2.5 0.9 5.4 29.2	4.4 1.5 2.3 8.2 32.5	2.1 1.2 0.5 3.8 16.9	2.3 1.3 0.8 4.5 22.3	2.4 1.4 0.9 4.6 23.7
unisys2	1.3 0.5 0.7 2.6 14.4	1.4 0.3 1.3 3.0 14.7	2.0 0.6 1.2 3.8 25.8	3.8 1.3 1.5 6.6 34.8	4.4 1.0 1.0 6.4 32.5	2.6 0.6 0.6 3.9 20.5	2.3 0.7 1.0 4.0 21.9	2.3 0.7 1.0 4.0 21.9
unisys3	0.6 0.1 0.5 1.3 6.8	1.5 0.4 1.1 3.0 15.3	1.9 0.5 1.2 3.6 24.2	3.5 1.5 1.9 7.0 32.6	4.9 0.7 0.7 6.2 28.6	3.1 0.9 0.3 4.3 20.5	2.3 0.6 0.9 3.9 19.7	2.3 0.6 0.9 3.9 19.7
Overall Totals	2.4 0.7 0.7 3.9 22.1	1.8 0.8 1.0 3.6 16.2	2.7 1.0 1.1 4.8 30.0	3.1 1.8 1.7 6.5 34.2	4.7 1.3 1.3 7.3 31.6	3.1 1.0 0.8 4.9 22.4		
Foreign System	2.6 0.8 0.7 4.2 24.0	1.9 0.8 1.0 3.7 16.8	2.7 0.9 1.1 4.7 29.7	3.1 1.8 1.7 6.5 34.2	4.7 1.3 1.3 7.3 31.6	3.4 0.9 0.9 5.2 24.0		

%Sub %Del %Ins %W.Err %Utt.Err

Matrix tabulation of results for the Dec93 ATIS SPREC Test Results, for the Class A+D Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present results for the entire Class A+D Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class A+D Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

Table 14 ATIS SPREC Results: Class (A+D) by Collection Site

Composite Report of All Significance Tests
For the Dec93 ATIS SPREC Class A+D Test Results Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy) Test	WI
McNemar (Sentence Error) Test	MN

	att2-a_d	bbn3-a_d	cmu2-a_d	crim3-a_d	mit_lcs2-a_d	sri3-a_d	sri4-a_d	unisys2-a_d	unisys3-a_d
att2-a_d		MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP cmu2-a_d SI cmu2-a_d WI cmu2-a_d MN cmu2-a_d	MP crim3-a_d SI crim3-a_d WI crim3-a_d MN crim3-a_d	MP mit_lcs2-a_d SI mit_lcs2-a_d WI mit_lcs2-a_d MN mit_lcs2-a_d	MP sri3-a_d SI sri3-a_d WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP unisys2-a_d SI unisys2-a_d WI unisys2-a_d MN unisys2-a_d	MP unisys3-a_d SI unisys3-a_d WI unisys3-a_d MN unisys3-a_d
bbn3-a_d			MP same SI same WI same MN same	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI same WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI same WI bbn3-a_d MN same
cmu2-a_d				MP cmu2-a_d SI cmu2-a_d WI cmu2-a_d MN cmu2-a_d	MP cmu2-a_d SI same WI same MN cmu2-a_d	MP cmu2-a_d SI same WI same MN cmu2-a_d	MP cmu2-a_d SI same WI same MN same	MP cmu2-a_d SI same WI same MN same	MP same SI same WI same MN same
crim3-a_d					MP mit_lcs2-a_d SI mit_lcs2-a_d WI mit_lcs2-a_d MN mit_lcs2-a_d	MP sri3-a_d SI sri3-a_d WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP unisys2-a_d SI unisys2-a_d WI unisys2-a_d MN unisys2-a_d	MP unisys3-a_d SI unisys3-a_d WI unisys3-a_d MN unisys3-a_d
mit_lcs2-a_d						MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP unisys3-a_d SI same WI same MN unisys3-a_d
sri3-a_d							MP same SI same WI same MN same	MP same SI same WI same MN same	MP unisys3-a_d SI same WI same MN unisys3-a_d
sri4-a_d								MP same SI same WI same MN same	MP same SI same WI same MN same
unisys2-a_d									MP same SI same WI same MN unisys3-a_d
unisys3-a_d									

Table 15 Significance Test Results: ATIS SPREC Systems

system	Class A+D 773 Utts. UW Err.	Class A 448 Utts. UW Err.	Class D 352 Utts. UW Err.
att1	10.2	7.4	14.2
bbn1	14.7	9.6	21.8
bbn2	22.4	16.1	31.1
cmu1	9.3	6.0	13.8
crim1	36.4	21.7	56.6
crim2	20.8	14.7	29.2
mit_lcs1	12.5	10.0	16.0
sri1	21.9	14.3	32.3
sri5 **	18.2	10.5	28.9
unisys1	43.1	28.6	63.1

Table 16 ATIS NL Test Results

Class (A+D) Set

Originating Site of Test Data

SYSTEMS	BBN 146	CMU 163	MIT 132	NIST-SRI 77	NIST-BBN 89	SRI 166	Overall Totals 773	Foreign Coll. Site Totals
att1	124 22 0 85 15 0 15.1	153 10 0 94 6 0 6.1	121 11 0 92 8 0 8.3	67 10 0 87 13 0 13.0	80 9 0 90 10 0 10.1	149 17 0 90 10 0 10.2	694 79 0 90 10 0 10.2	694 79 0 90 10 0 10.2
bbn1	124 21 1 85 14 1 15.1	141 22 0 87 13 0 13.5	117 15 0 89 11 0 11.4	57 20 0 74 26 0 26.0	82 7 0 92 8 0 7.9	138 28 0 83 17 0 16.9	659 113 1 85 15 0 14.7	535 92 0 85 15 0 14.7
bbn2	104 41 1 71 28 1 28.8	127 36 0 78 22 0 22.1	112 20 0 85 15 0 15.2	54 23 0 70 30 0 29.9	69 20 0 78 22 0 22.5	134 32 0 81 19 0 19.3	600 172 1 78 22 0 22.4	496 131 0 79 21 0 20.9
cmu1	128 18 0 88 12 0 12.3	153 10 0 94 6 0 6.1	120 12 0 91 9 0 9.1	67 10 0 87 13 0 13.0	80 9 0 90 10 0 10.1	153 13 0 92 8 0 7.8	701 72 0 91 9 0 9.3	548 62 0 90 10 0 10.2
crim1	76 61 9 52 42 6 47.9	114 31 18 70 19 11 30.1	88 36 8 67 27 6 33.3	40 32 5 52 42 6 48.1	79 10 0 89 11 0 11.2	95 57 14 57 34 8 42.8	492 227 54 64 29 7 36.4	492 227 54 64 29 7 36.4
crim2	112 33 1 77 23 1 23.3	133 27 3 82 17 2 18.4	109 23 0 83 17 0 17.4	57 17 3 74 22 4 26.0	76 12 1 85 13 1 14.6	125 41 0 75 25 0 24.7	612 153 8 79 20 1 20.8	612 153 8 79 20 1 20.8
mit_lcs1	111 35 0 76 24 0 24.0	150 13 0 92 8 0 8.0	120 12 0 91 9 0 9.1	62 15 0 81 19 0 19.5	83 6 0 93 7 0 6.7	150 16 0 90 10 0 9.6	676 97 0 87 13 0 12.5	556 85 0 87 13 0 13.3
sri1	103 17 26 71 12 18 29.5	130 17 16 80 10 10 20.2	111 8 13 84 6 10 15.9	54 21 2 70 27 3 29.9	75 14 0 84 16 0 15.7	131 30 5 79 18 3 21.1	604 107 62 78 14 8 21.9	473 77 57 78 13 9 22.1
sri5 **	113 30 3 77 21 2 22.6	142 17 4 87 10 2 12.9	117 14 1 89 11 1 11.4	54 21 2 70 27 3 29.9	75 14 0 84 16 0 15.7	131 30 5 79 18 3 21.1	632 126 15 82 16 2 18.2	501 96 10 83 16 2 17.5
unisys1	76 31 39 52 21 27 47.9	88 33 42 54 20 26 46.0	91 26 15 69 20 11 31.1	40 24 13 52 31 17 48.1	49 27 13 55 30 15 44.9	96 20 50 58 12 30 42.2	440 161 172 57 21 22 43.1	440 161 172 57 21 22 43.1
Overall Totals	1071 309 80 73 21 5 26.6	1331 216 83 82 13 5 18.3	1106 177 37 84 13 3 16.2	552 193 25 72 25 3 28.3	748 128 14 84 14 2 16.0	1302 284 74 78 17 4 21.6		
Foreign System Totals	843 247 78 72 21 7 27.8	1178 206 83 80 14 6 19.7	986 165 37 83 14 3 17.0	552 193 25 72 25 3 28.3	748 128 14 84 14 2 16.0	1040 224 64 78 17 5 21.7		

Legend:

#T	#F	#NA
%T	%F	%NA
% Un-Weighted Err		

Matrix tabulation of results for the Dec 93 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion, for the Class (A+D) Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of evaluable utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present results for the entire Class (A+D) Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class (A+D) Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

** Late and for a debugged system.

Table 17 ATIS NL Results: Class (A+D) by Collection Site

	Class A+D 773 Utts	Class A 448 Utts.	Class D 352 Utts.
system	UW Err.	UW Err.	UW Err.
att1	24.6	22.1	28.0
bbn1	17.5	13.8	22.5
cmu1	13.2	8.9	19.1
crim1	43.3	28.6	63.7
crim2	28.2	23.7	34.5
mit_lcs1	14.2	11.8	17.5
sri1	24.8	16.5	36.3
sri2	25.4	18.5	34.8
sri5 **	20.7	14.1	29.8
sri6 **	21.2	13.8	31.4
unisys1	46.8	33.5	65.2

Table 18 ATIS SLS Test Results

Class (A+D) Set
Originating Site of Test Data

SYSTEMS	BBN 146	CMU 163	MIT 132	NIST-SRI 77	NIST-BBN 89	SRI 166	Overall Totals 773	Foreign Coll. Site Totals
att1	106 40 0 73 27 0 27.4	138 25 0 85 15 0 15.3	100 32 0 76 24 0 24.2	58 19 0 75 25 0 24.7	72 17 0 81 19 0 19.1	109 57 0 66 34 0 34.3	583 190 0 75 25 0 24.6	583 190 0 75 25 0 24.6
bbn1	121 24 1 83 16 1 17.1	128 35 0 79 21 0 21.5	117 15 0 89 11 0 11.4	61 16 0 79 21 0 20.8	75 14 0 84 16 0 15.7	136 30 0 82 18 0 18.1	638 134 1 83 17 0 17.5	517 110 0 82 18 0 17.5
cmu1	127 19 0 87 13 0 13.0	152 11 0 93 7 0 6.7	114 18 0 86 14 0 13.6	64 13 0 83 17 0 16.9	76 13 0 85 15 0 14.6	138 28 0 83 17 0 16.9	671 102 0 87 13 0 13.2	519 91 0 85 15 0 14.9
crim1	73 65 8 50 45 5 50.0	100 44 19 61 27 12 38.7	82 42 8 62 32 6 37.9	38 35 4 49 45 5 50.6	66 21 2 74 24 2 25.8	79 72 15 48 43 9 52.4	438 279 56 57 36 7 43.3	438 279 56 57 36 7 43.3
crim2	104 40 2 71 27 1 28.8	121 40 2 74 25 1 25.8	99 33 0 75 25 0 25.0	55 20 2 71 26 3 28.6	69 18 2 78 20 2 22.5	107 58 1 64 35 1 35.5	555 209 9 72 27 1 28.2	555 209 9 72 27 1 28.2
mit_lcs1	110 36 0 75 25 0 24.7	148 15 0 91 9 0 9.2	116 16 0 88 12 0 12.1	63 14 0 82 18 0 18.2	81 8 0 91 9 0 9.0	145 21 0 87 13 0 12.7	663 110 0 86 14 0 14.2	547 94 0 85 15 0 14.7
sri1	94 19 33 64 13 23 35.6	140 20 3 86 12 2 14.1	99 11 22 75 8 17 25.0	46 16 15 60 21 19 40.3	68 20 1 76 22 1 23.6	134 28 4 81 17 2 19.3	581 114 78 75 15 10 24.8	447 86 74 74 14 12 26.4
sri2	100 15 31 68 10 21 31.5	121 27 15 74 17 9 25.8	103 12 17 78 9 13 22.0	53 23 1 69 30 1 31.2	67 20 2 75 22 2 24.7	133 29 4 80 17 2 19.9	577 126 70 75 16 9 25.4	444 97 66 73 16 11 26.9
sri5 **	105 38 3 72 26 2 28.1	140 20 3 86 12 2 14.1	112 19 1 85 14 1 15.2	54. 21 2 70 27 3 29.9	68 20 1 76 22 1 23.6	134 28 4 81 17 2 19.3	613 146 14 79 19 2 20.7	479 118 10 79 19 2 21.1
sri6 **	111 32 3 76 22 2 24.0	133 27 3 82 17 2 18.4	112 19 1 85 14 1 15.2	53 23 1 69 30 1 31.2	67 20 2 75 22 2 24.7	133 29 4 80 17 2 19.9	609 150 14 79 19 2 21.2	476 121 10 78 20 2 21.6
unisys1	72 36 38 49 25 26 50.7	88 31 44 54 19 27 46.0	84 33 15 64 25 11 36.4	33 29 15 43 38 19 57.1	49 29 11 55 33 12 44.9	85 30 51 51 18 31 48.8	411 188 174 53 24 23 46.8	411 188 174 53 24 23 46.8
Overall Totals	1123 364 119 70 23 7 30.1	1409 295 89 79 16 5 21.4	1138 250 64 78 17 4 21.6	578 229 40 68 27 5 31.8	758 200 21 77 20 2 22.6	1333 410 83 73 22 5 27.0		
Foreign System Totals	1002 340 118 69 23 8 31.4	1257 284 89 77 17 5 22.9	1022 234 64 77 18 5 22.6	578 229 40 68 27 5 31.8	758 200 21 77 20 2 22.6	799 296 67 69 25 6 31.2		

Legend:

#T	#F	#NA
%T	%F	%NA
% Un-Weighted Err		

Matrix tabulation of results for the Dec 93 ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion, for the Class (A+D) Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of evaluable utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present results for the entire Class (A+D) Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class (A+D) Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

** Late and for a debugged system.

Table 19 ATIS SLS Results: Class (A+D) by Collection Site