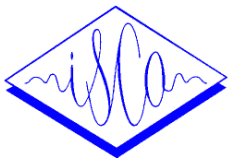

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

BENCHMARK TESTS FOR DARPA RESOURCE MANAGEMENT DATABASE PERFORMANCE EVALUATIONS

David S. Pallett

Room A216 Technology Building
National Institute for Standards and Technology
Gaithersburg, Maryland 20899

ABSTRACT

The implementation of benchmark test procedures making use of the DARPA Resource Management speech database has made it possible to monitor progress of the development of speech recognition technology within the DARPA Speech Research community. This paper outlines a number of considerations that must be taken in implementing benchmark tests such as these.

1. INTRODUCTION

Much of this paper consists of an adaptation of material appearing in an identically titled paper appearing in the Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing (Pallett, 1989). The reader is referred to that paper for additional details and for additional references. Recent experience (since submission of the text of that paper) is outlined in this paper.

2. COMMON TASK

For the DARPA Speech Recognition Program, it was agreed that the initial task to be pursued was speech recognition, rather than, (at that time) speech understanding. Thus the task to be performed by the systems was limited to orthographic transcriptions of the spoken utterances. No information was required at the sub-word level. Only one hypothesis, or transcription, was required and to be scored.

Only one task domain and speech corpus was to be used for the benchmark tests, the DARPA Resource Management Speech Database.

Agreement on these issues greatly simplified the task of developing and implementing benchmark tests.

3. SELECTION OF TEST MATERIAL

The National Institute of Standards and Technology (NIST: formerly NBS) selected a number of test sets for the benchmark tests. Test sets were selected for use at CMU and BBN in March of 1987, and in October 1987. Additional material was designated for use in May-June 1988, in February 1989, and for October 1989. By February 1989, test results had been reported to NIST by the following organizations: AT&T Bell Labs, BBN, CMU, MIT Lincoln Lab, MIT Laboratory for Computer Science, and SRI.

Although effort was expended to randomize the selection of the test material, there is some evidence that the test sets were slightly different in the degree of difficulty. Some of this is attributed to within-session effects, and some to between session effects for the speaker-dependent material. These findings underscore the desirability of randomizing the selection of test material.

4. SYSTEM TRAINING

It was necessary to specify the allowable use of the speech corpora material for system training.

For the speaker-dependent systems, a total of 570 read sentences (for each speaker) was to be used for system training. For the speaker-independent systems, it was necessary to designate two system training conditions, since there was a misunderstanding about what could (and could not) be used. By February 1989, agreement had been reached on the use of two "official" training sets: (1) The "Standard Condition", consisting of 72 speakers, and (2) The "Augmented (or extra training data) Condition, consisting of the Standard Condition augmented by another set of 37 speakers, for a total of 109 speakers in all.

5. CONVENTIONS

In order to implement uniform scoring procedures at the word level, operating on the orthographic transcriptions, it was necessary to specify a "standardized normal orthographic representation" (SNOR) for each word in the lexicon. SNOR versions of the texts that the subjects read for the resource management database were also generated. Additional conventions were developed defining the format of the system output for scoring by the scoring software.

6. CONSTRAINING GRAMMARS

Two conditions were defined for the constraining grammars used for these tests: (1) a "no grammar" case, in which each word is equally probable, regardless of preceding words, and (2) a non-probabilistic "word-pair" grammar, that specifies the set of permissible words that may follow each word.

7. SCORING SOFTWARE

Scoring software was developed at NIST using input from BBN, MIT/Lincoln Laboratory and TI. Using the same set of scoring software tools at each site has made it possible to compare system performance without the complication of using conflicting measures.

7.1 String Matching

Fundamental to all of the scoring software was some form of dynamic-programming string alignment procedure. The string alignment software aligns two strings of words: the reference SNOR transcription of the test material, and the hypothesized transcription that is the output of the speech recognizer. The simplest approach makes no use of sub-word information. It treats all words as merely as symbols, and all confusion pairs, insertions and deletions are penalized equally when implementing the string alignment.

7.1.1 Dynamic Programming Symbol String alignment

In the present string alignment procedure, the algorithm computes the lowest cost alignments between two word strings, given the constraints that exact matches incur no penalty, deletion and insertion penalties are equal, and the sum of the penalties for one deletion and one insertion is greater than one substitution penalty. We have corresponded with Hunt (Hunt, 1988) about some of the limitations of this approach, and feel that the degree of imprecision inherent to this approach is not severe, particularly for high-performance systems.

7.1.2 Phonetically Motivated Approaches

In another procedure for string alignment, the word strings are converted to phone strings using any of several approaches. The phone strings are then aligned using any of several phone-to-phone substitution, insertion or deletion weighting schemes. Finally, the word boundaries are inferred from the aligned phone strings. Picone and Doddington (Picone et al. 1989) have shared with us one such scheme, and in general it appears to yield better diagnostics than the simpler symbol string matching

alignment procedure. At NIST, we are presently developing an alternative procedure that uses phonetic information for string alignment, and we may incorporate that procedure into a revised scoring software package.

7.2 Error Taxonomy

Scoring at the word level generates statistics on the percentage of words in the reference strings that were correctly recognized, and for errors classified as substitutions, deletions and insertions. Reports of benchmark tests frequently report on the total (unweighted) percentage of errors, and on the "word accuracy", which is defined as 100% minus the total error percentage.

Although presentations including benchmark test results frequently present only one of these error statistics (and by implication suggest that performance can be summarized in the one statistic) it is preferable that the several statistics be presented, since different sites may to optimize their systems differently (e.g., with different insertion error rates). Furthermore, as Hunt correctly points out, the percentage of correctly recognized words may be a more reliable indicator of performance than measures of the individual errors.

7.3 Special Provisions

There are special provisions in the scoring software to permit "forgiving" substitution errors involving homophone errors for the case of no grammar, and for computing separate sets of statistics for mono-syllabic and poly-syllabic lexical items.

7.4 Documentation

The scoring software provides a uniform form of summary tabulation for the results that makes it easy to comprehend the results of the test. In the tests implemented within the DARPA program for the "official benchmark test", NIST has made use of this capability to prepare and distribute summaries of each site's results.

8. STATISTICAL CONSIDERATIONS

As Gillick and Cox note, "In the development of speech recognition algorithms, it is important to know whether any apparent difference in performance of algorithms is statistically significant, yet this issue is almost always overlooked" (Gillick and Cox 1989). Within the DARPA program, it became evident from study of the benchmark test results that a number of sites had high-performance systems, and the question arose if the differences in performance were significant. To address this issue, NIST has implemented the suggestions of Gillick and Cox and have also implemented a two way analysis of variance for the results of the February 1989 tests.

8.1 General Concerns

The case of testing the statistical significance of benchmark test results for the recognition of continuous speech is particularly complex. Because of coarticulation across word boundaries, the errors that recognizers make can not be expected to be totally independent, because of errors at the spoken sub-word level that are not independent. Further, the design of some recognizers is such that, because of the internal modelling that may be used, errors at the word level may influence the recognition of succeeding words. Thus the set of assumptions that can be used in testing statistical significance is always open to question. This is an area that needs further study.

8.2 McNemar's Test

McNemar's test can be thought of as a test that focuses on the errors that are unique to each of two systems, treatments, or algorithms, placing less emphasis on errors or correct recognitions that occur in

common. Gillick and Cox (1989) suggest that for the case of continuous speech, errors are apt to be highly inter-dependent, and implementing McNemar's test at the word level would be inappropriate. They suggest an alternative implementation at the sentence level, where a the system output for each sentence is considered either correct (i.e., with no errors) or incorrect. We have implemented this procedure at NIST in analyzing the results of the February 1989 benchmark tests.

8.3 Matched-Pairs Test

An alternative to the NcNemar test is the matched-pairs test described by Gillick and Cox (1989). In our present implementation, a string alignment procedure is used to identify segments of each sentence that contain errors, but that are bounded on the left by either the first word of a sentence utterance, or by one (or more) correctly recognized words. The right boundary of the segment is, correspondingly, either the last word in the sentence utterance or one (or more) correctly recognized words. Corresponding segments of the sentence hypotheses of two algorithms or systems are then compared. For each of the two corresponding segments, a count of total errors is made, and the difference is computed. A null hypothesis test can then be applied to measures of the observed average difference in the number of errors in a segment. In our preliminary implementation of this test to our data, this appears to be somewhat more sensitive than the McNemar test at the sentence level.

ACKNOWLEDGEMENTS

In addition to those credited as contributing significantly to the work reported in (Pallett 1989), more recently, at NIST, Jon Fiscus and Roy Gengel have contributed to, and implemented the extensions involving the phonetically motivated string alignment approach and to the incorporation of statistical measures.

Discussions with Larry Gillick at Dragon Systems have been very instructive and are greatly appreciated.

REFERENCES

Gillick, L. and Cox, S. (1989) Some Statistical Issues in the Comparison of Speech Recognition Algorithms, Proceedings of ICASSP '89, Paper S10b.5, pp. 532-535.

Hunt, M.R. (1988) Evaluating the Performance of Connected Word Speech Recognition Systems, Proceedings of ICASSP '88, Paper S.10.13, pp. 457-460.

Pallett, D.S. (1989) Benchmark Tests for DARPA Resource Management Database Performance Evaluations, Proceedings of ICASSP '89, Paper S10b.6, pp. 536-539.

Picone, J., Doddington, G.R. and Pallett, D.S. (1989) Phone-Mediated Word Alignment for Speech Recognition Evaluation, submitted to the IEEE Transactions on Acoustics, Speech and Signal Processing.