# The Hub and Spoke Paradigm for CSR Evaluation

*Francis Kubala* *BBN Systems and Technologies*
*Jerome Bellegarda* *IBM T. J. Watson Research Center*
*Jordan Cohen* *Institute for Defense Analyses*
*David Pallett* *National Institute of Standards and Technology*
*Doug Paul* *MIT Lincoln Laboratory*
*Mike Phillips* *MIT Laboratory for Computer Science*
*Raja Rajasekaran* *Texas Instruments*
*Fred Richardson* *Boston University*
*Michael Riley* *AT&T Bell Laboratories*
*Roni Rosenfeld* *Carnegie Mellon University*
*Bob Roth* *Dragon Systems, Inc.*
*Mitch Weintraub* *SRI International*

1993 Members of the CSR Corpus Coordinating Committee (CCCC)
e-mail: cccc@bbn.com

## ABSTRACT

In this paper, we introduce the new paradigm used in the most recent ARPA-sponsored Continuous Speech Recognition (CSR) evaluation and then discuss the important features of the test design.

The 1993 CSR evaluation was organized in a novel fashion in an attempt to accomodate research over a broad variety of important problems in CSR while maintaining a clear program-wide research focus. Furthermore, each test component in the evaluation was designed as an experiment to extract as much information as possible from the results.

The evaluation was centered around a large vocabulary speaker-independent (SI) baseline test, which was required of every participating site. This test was dubbed the 'Hub' since it was common to all sites and formed the basis for controlled *inter*-system comparisons.

The Hub test was augmented with a variety of problem-specific optional tests designed to explore a variety of important problems in CSR, mostly involving some kind of mismatch between the training and test conditions. These tests were known as the 'Spokes' since they all could be informatively compared to the Hub, but were otherwise independent.

In the first trial of this evaluation paradigm in November, 1993, 11 research groups participated, yielding a rich array of comparative and contrastive results, all calibrated to the current state of the art in large vocabulary CSR.

## 1. Introduction

Since 1986, ARPA has sponsored periodic formal evaluations of CSR technology. From the beginning, these evaluations were distinguished by a well-defined test that was required of all participants within a specified window of time. Most importantly, the test definition remained stable over several years so that sustained effort could be made toward improved performance and so that any improvement made over time could be demonstrated convincingly. These were important features of a series of evaluations based on the well-known Resource Management (RM) corpus. Those evaluations were highly regarded for the competitive stimulus they produced, resulting in the rapid assimilation of new techniques across the CSR community worldwide.

When the ARPA CSR community began designing a testbed corpus for large-vocabulary recognition to replace RM, one of the shortcomings addressed was its lack of support for the variety of important research interests that existed within the community at the time. Active research was already underway in adaptation to speaker, domain, dialect, and in compensation for mismatch in microphone, environment, and speaking style, but none of this was supported by the RM corpus.

The ARPA-chartered CSR Corpus Coordinating Committee (CCCC) was given the task of defining a corpus and specifying an evaluation scheme that would take advantage of this diversity and drive it to produce enabling technology for eventual application to real-world CSR problems.

The Hub and Spoke evaluation paradigm was conceived to accomodate the research requirements of this diverse community and produce convincing demonstrations of technological capability. Tests were defined, to exercise the primary interests of all participants, and to include important comparisons needed to make informed descisions about the efficacy of a particular algorithm or general approach. At the same time, the evaluation preserved the important controlled baseline test, characteristic of past ARPA-sponsored evaluations, that permitted direct comparison of CSR technology across different systems.

In the next section, we describe the general design of the Hub and Spoke evaluation paradigm. Each component test in the 1993 evaluation is then described in detail in section 3.

## 2. The Hub and Spoke Evaluation Paradigm

The *Hub and Spoke* appelation is intended to characterize the organization of the evaluation as a suite of fairly independent tests (the Spokes) coupled to a central test (the Hub) in some informative fashion. The Hub test is further distinguished by being an abstract representation of a fundamentally important problem in CSR and by being the only test required of all participants in the evaluation. It forms the basis for all informative *inter*-system comparisons.

The Spoke tests, on the other hand, are abstractions of problems of somewhat less central importance in CSR and evaluation on them is optional. They are the research sandbox, if you will, where new problems and methods can be introduced and evaluated speculatively, without requiring agreement of the entire ARPA CSR community. They are specifically designed to permit *intra*-system comparisons of algorithms and methods for problems that often involve a mismatch between training and test data. The Spoke tests can all be informatively compared to the Hub test but they are otherwise independent.

Every Hub or Spoke test was structured to produce a primary result for each system evaluated as well as one or more contrastive conditions designed to measure the effect of an algorithm or approach on the problem under study. For instance, the primary result might have featured a noise compensation algorithm while a contrastive test, on the same data, might have required that the compensation be disabled. Comparing these two results demonstrates the efficacy of the compensation.

The primary conditions are designated the P0 condition. In general, they are unconstrained with respect to the lexicon and acoustic or language model (LM) training allowed. The contrastive condition(s) are designated as CX (X = 1,2,..). The first contrast test (designated C1) normally specifies that an adaptive or unconstrained feature of the primary test be disabled or constrained so that the effect of the primary feature can be measured in isolation. This contrast is usually required. Additional contrastive tests (either required or optional) may be specified to calibrate the data or evaluate the featured algorithm on additional data.

The C1 contrast condition in the Hub test has special importance in the overall evaluation design. This condition specifies exactly the acoustic training data allowed and the precise LM to be used. These easily-varied parameters of the test are held fixed in this one condition in order to focus attention solely upon the acoustic modeling power of each system tested. Although fixed, the amount of training data permitted and the quality of the LM used in the C1 test are near state-of-the-art. This controlled test therefore establishes a convincing baseline which allows a direct comparison between all systems in the evaluation.

The P0 primary condition in the Hub also occupies an exalted position within the evaluation framework. It is designed to test the current capability on a central problem in CSR. In a single number, results from this test quantify the meaning of state-of-the-art in large-vocabulary continuous speech recognition.

**Terminology**

Several useful terms are defined here that describe important features of the evaluation.

A *session* implies that the speaker, microphone, and acoustic environment all remain constant for a group of utterances.

A *static SI* test does not offer session boundaries or utterance order as side information to the system, and therefore implies that the speaker, microphone, and environment may change from utterance to utterance. Functionally, it implies that each utterance must be recognized independently of all others, yielding the same answers for any utterance order (or the same expectation, in the case of a non-deterministic recognizer).

*Unsupervised incremental adaptation* means that the system is allowed to use any information it can extract from test data that it has already recognized. It implies that session boundaries and utterance order are known to the system as side information.

*Supervised incremental adaptation* means that the correct transcription is made available to the system after each utterance has been recognized. It also implies that session boundaries and utterance order are known to the system as side information. This recognition mode models the scenario in which the user incrementally corrects the system response after each utterance.

## 3. The 1993 Hub and Spoke Evaluation

The Hub and Spoke evaluation paradigm was first used for ARPA CSR evaluation in November, 1993. The entire test suite for this evaluation consisted of 2 Hub tests and 9 Spoke tests. Each of these tests is described in detail below.

| *Designation* | *Vocabulary* |
|---|---|
| *THE HUB* | |
| H1. Read WSJ Baseline | 64K |
| H2. 5K-Word Read WSJ Baseline | 5K |
| *THE SPOKES* | |
| S1. Language Model Adaptation | unlimited |
| S2. Domain-Independence | unlimited |
| S3. SI Recognition Outliers | 5K |
| S4. Incremental Speaker Adaptation | 5K |
| S5. Microphone-Independence | 5K |
| S6. Known Alternate Microphone | 5K |
| S7. Noisy Environments | 5K |
| S8. Calibrated Noise Sources | 5K |
| S9. Spontaneous WSJ Dictation | unlimited |

The abstract problem represented by all the tests in the 1993 evaluation was the dictation of news stories, with an emphasis on financial news stories. Most of the tests in the 1993 evaluation used speech data from subjects reading selected articles from the Wall Street Journal. The prompting texts for the WSJ-based tests came from the pre-defined evaluation test text pools specified in the WSJ0 corpus [3] which consists of articles from the Wall Street Journal published during the years 1987-1989.

Typical tests used 10 subjects reading 20-40 sentences each. Each test had equal numbers of male and female subjects. The Sennheiser HMD-410 close-talking, noise-canceling microphone was the primary one used.

Unless otherwise noted, the default side information given to the system was as follows. Speaking style, general environment conditions (quiet or noisy), and microphone identity were known. Speaker gender, specific environment conditions (room identity), session boundaries and utterance order were unknown unless noted otherwise. Collectively, these defaults imply that static SI conditions were the default.

### 3.1. The 1993 Hub

The Hub for 1993 was split into two tests differing in vocabulary size (64K and 5K-words). The smaller test was included to provide a computationally tractable test for sites that were not prepared at the time to handle the larger vocabulary at the time.

**H1. Read WSJ Baseline**

The paramount Hub test (H1) was designed to measure state-of-the-art performance on a large-vocabulary SI test, using clean test data well-matched to the training data. The prompting texts for the H1 test came from the pre-defined 64K-word WSJ0 text pools. These texts excluded paragraphs that contained words outside the 64K most frequent from the WSJ0 corpus.

The primary H1 (H1-P0) test allowed any language model (LM) or acoustic training data to be used. In addition, the temporal order of the utterances and the location of subject-session boundaries in the utterance sequence was given to encourage the use of unsupervised incremental adaptation techniques.

To permit direct comparisons of acoustic modeling technology between different systems, the H1 test contained a required contrastive test (H1-C1) that controlled the amount of training data and specified the LM statistics. This contrast was run as a static SI test, so utterance order and session boundaries were not given to the system.

For H1-C1, the acoustic training data was limited to 37.2K utterances (about 62 hours of speech) drawn from one of two segments of the combined WSJ0 and WSJ1 corpora. One segment was made up of speech data from 284 subjects (SI-284) who produced 100-150 utterances each. The other segment had 37 subjects (SI-37) who produced either 600 or 1200 utterances each. Evaluating sites were free to choose either acoustic training corpus.

The common required LM specified for the H1-C1 test was produced by Doug Paul at MIT Lincoln Laboratory. It was a 3-gram backoff LM estimated from approximately 35M words of text in the 1987-89 WSJ0 text corpus. Its lexicon was defined as the 20K most frequent words in the corpus, hence, the test contained some words outside the vocabulary. For the closed-vocabulary version of the 20K trigram LM, the perplexity is about 160.

An optional contrast, H1-C2, was specified with a companion 20K-word bigram LM produced by Doug Paul. All other conditions were identical to H1-C1.

**H2. 5K-Word Read WSJ Baseline**

The smaller 5K Hub test used prompting texts from the 5K-word text pools specified in the WSJ0 corpus. These articles were filtered to discard paragraphs with more than one word outside of the 5K most frequent words in the corpus.

Similar to the H1-P0 test, the primary H2 test (H2-P0) allowed any language model (LM) or acoustic training data to be used and also allowed unsupervised incremental adaptation.

The required H2-C1 was scaled down, however, to reduce the computational burden of participation. The acoustic training data was limited to 7.2K utterances (about 12 hours of speech) drawn from either the short-term or long-term subject segments of the combined WSJ0 and WSJ1 corpora. Here, the short-term subjects numbered 84 (SI-84), compared to 12 (SI-12) subjects for the long-term segment.

A common 5K-word bigram LM, produced at MIT Lincoln Laboratory was required for H2-C1. This LM was nominally a closed-vocabulary grammar. The lexicon was constructed by including all the words from the test truth texts and then adding words from WSJ0 word-frequency-list until 5K words were accumulated. Due to subject variability in reading the prompting texts, a few words were produced that were outside the specified vocabulary. The perplexity of the standard 5K closed-vocabulary LM is about 80 for the bigram and 45 for the trigram.

## 3.2. The 1993 Spokes

There were 9 Spoke tests in the 1993 evaluation that were designed to support the major interests of the participating sites at the time. Most of them are designed to study problems involving a mismatch between the training and test data.

Spokes S1 and S2 supported problems in LM adaptation primarily. S3 and S4 were targeted at speaker adaptation methods. Adaptation to microphone was the focus of Spokes S5 and S6. Ambient noise was considered in S7 and S8. Spoke S9 looked at data from a potential application for large-vocabulary CSR – spontaneous dictation of news stories from print-media journalists.

All Spokes except S1, S2, and S9 used read-speech from the WSJ0 5K-word prompting texts. The channel and noise compensation Spokes, S5-S8, used data from a variety of secondary microphones. All other Spokes used data from the Sennheiser microphone.

In each Spoke below, the primary test represents the abstract problem of interest. The system is generally the least constrained for the primary condition but run on data that is somehow mismatched to the training data. The contrastive tests then attempt to expose information by constraining some feature of the primary system or data. Most often the first contrast (C1) constrains the *system* to show the efficacy of an algorithm used in the primary test. Other contrasts, which may vary in number from Spoke to Spoke, will often be run on the matched *data* to calibrate the problem and estimate upper-bound performance.

Since the purpose of the Spoke tests is to calibrate the effectiveness of an algorithm or approach within a single system, sites are free to choose their system parameters for the primary condition as they see fit. That means, however, that direct comparisons between systems cannot be made without suitable caution. Unless the system details are well understood, the reader could unknowingly end up comparing a system using a bigram LM to one using a trigram, for instance, and thereby draw a completely inappropriate conclusion. In general it is advisable to make only *intra*-system comparisons between results in all Spoke tests.

**S1. Language Model Adaptation**

This Spoke was concerned with the problem of within-domain sub-language adaptation. The data was read-speech prompted from unfiltered texts of 1990 WSJ publications. Articles were selected with a minimum of 20 sentences.

For the primary test (S1-P0) the vocabulary was closed and the recognition mode was specified as supervised incremental adaptation, so a system was allowed to know all words that could occur and was given the correct answer after each utterance was recognized. The sequential order of the utterances and the article boundaries were known. The LM training was restricted to the 1987-1989 WSJ0 texts so that it predated any of the test data.

The required contrast, S1-C1, repeated the same test with the LM adaptation disabled. An optional contrast, S1-C2, specified unsupervised incremental LM adaptation (the vocabulary was closed but the correct answer was not given).

For reporting results, the test data was partitioned into four parts that were distinguished by their position within the articles. Separate word error rates were given for utterances 1-5, 6-10, 11-15, and 16 or greater in a given article. This was done to observe the effect of the LM adaptation as a function of the amount of context available to it.

### S2. Domain-Independence

The purpose of this Spoke was to evaluate techniques for dealing with a newspaper domain different from the training. The data was prompted from articles drawn from the San Jose Mercury (SJM) newspaper.

The primary test allowed any acoustic or LM training with the restrictions that no training be used from the San Jose Mercury itself and that no use could be made of knowledge of the paper's identity. Two required contrastive tests then calibrated the S2-P0 system on the WSJ-based H1 data, and conversely, the H1-P0 system on the SJM-based S2 data.

As it happened, no site evaluated on this Spoke in 1993.

### S3. SI Recognition Outliers

This Spoke was designed to study speaker adaptation for non-native speakers of American English, a group for which recognition performance is often very degraded. The goal of the test was to reduce this degradation by using limited enrollment data from each of these speakers to adapt the system to them.

The data was read-speech from the 5K-word WSJ texts similar to the H2 data. It was produced by ten subjects whose first languages included, French, Spanish, German, Danish, Japanese, Hebrew, and British English.

The primary test allowed use of 40 utterances of rapid-enrollment data collected from each of the test speakers. The required S3-C1 contrast was then run with the adaptation disabled to measure its effectiveness. A second required contrast test ran the S3-P0 adaptive system on the H2 data to see the effect of the adaptation on native speakers of the language. All tests were static SI recognition tests.

### S4. Incremental Speaker Adaptation

Spoke S4 was directed specifically toward incremental speaker adaptation for native speakers. The goal was to improve on the baseline SI performance by adapting to each test speaker individually.

The 5K-word WSJ read data for this Spoke differed from the H2 data only in that each speaker produced 100 utterances so that the convergence of the adaptation algorithms could be observed over a longer session from each speaker.

The primary test specified unsupervised incremental adaptation. The required S4-C1 contrast was run with the adaptation disabled. An optional test, S4-C2, allowed supervised incremental adaptation for comparison with the primary test.

For reporting results, the test data was partitioned into four parts that were distinguished by their position within the sessions. Separate word error rates were given for utterances 1-25, 26-50, 51-75, and 76 or greater within a given session. This was done to observe the effect of the speaker adaptation as a function of the amount of context available to it.

To measure the cost of the adaptation relative to the recognition, the ratio of the total runtime of S4-P0 to S4-C1 was also given as an auxilliary performance measure.

### S5. Microphone-Independence

Spoke S5 exposed the system to a variety (10) of unknown microphones. Only unsupervised compensation algorithms were allowed for this Spoke. It was a static SI test as well, so that each utterance had to be considered in isolation.

The data was the same speech as the H2 data (5K-word WSJ), but collected in stereo recordings through the various secondary microphones. The selection of microphones included telephone (handset and speakerphone), lapel, hand-held, stand-mounted, and monitor-mounted ones. They included professional quality microphones as well as consumer grade devices.

The primary test demonstrated the unsupervised compensation algorithm. The required S5-C1 test ran with the compensation disabled to show the degradation due to the mismatched channel. This condition calibrates the mismatched channel problem.

The S5-C2 test required the S5-P0 system (compensation enabled) to be run on the matching stereo channel from the Sennheiser microphone to observe the effect of the compensation on data that was matched to the training. An optional S5-C3 contrast calibrated the S5-C1 system (compensation disabled) on the stereo Sennheiser data to set the upper-bound matched-channel performance level.

These 4 conditions define the space of the mismatched channel problem completely. The ideal result is that the compensated mismatched condition (S5-P0) works as well as the uncompensated matched condition (S5-C3), and that the compensation doesn't hurt when used on the matched data (S3-C2).

### S6. Known Alternate Microphone

Spoke S6 exposed the system to two microphones whose identities were known, but that differed from the microphone used in the training. One was a telephone handset and the other was a high-quality directional stand-mounted microphone placed about 18 inches from the subject's mouth. The goal of the test was to make the performance of the mismatched (but known) channel the same as the matched channel data.

The test data was produced from the 5K-word WSJ prompts. The test mode was static SI recognition. Simultaneous stereo recordings were also made through the Sennheiser microphone to calibrate the channel mismatch.

To facilitate adaptation to the known microphone, an additional set of stereo recordings were made for both microphone types paired with the Sennheiser from 10 training speakers. This adaptation data was the only data allowed from the target microphones.

The primary test demonstrated an adaptation algorithm and allowed the use of the stereo microphone-adaptation data. There were two required contrasts. S6-C1 ran with the adaptation disabled for comparison to the primary test. S6-C2 ran the S6-C1 system (adaptation disabled) on stereo Sennheiser data to set the baseline performance for the matched-channel condition. Separate error rates are reported for the two microphones.

The telephone data was routed through the local Palo Alto digital network, but sampled on wide-band analog lines in the internal phone system at SRI. Just prior to the 1993 evaluation, problems with this collection procedure were discovered that resulted in marked differences between the microphone-adaptation data and the development test data for this Spoke. Any algorithms that used the adaptation data may have been adversely affected by this difference.

### S7. Noisy Environments

This Spoke featured the same two microphones used in S6 but placed them in noisy environments with a background A-weighted noise level ranging over 55-68 dB. Sennheiser data from stereo recordings was also produced for calibration tests. The goal of the test was to minimize the difference in performance between the noise-cancelling Sennheiser microphone and the alternate microphones (telephone handset and far-field stand-mounted microphone).

The two environments sampled were large rooms housing an open computer laboratory and a mechanical equipment laboratory. Noise in the computer lab came from the surrounding equipment and normal human traffic around the lab. In the mechanical lab, the noises were generated by parcel sorting equipment.

The test data was from the 5K-word WSJ prompts. The test mode was static SI recognition. As in Spoke S6, use of the stereo microphone-adaptation data was permitted.

The primary test demonstrated a noise compensation algorithm and the required S7-C1 contrast ran with the compensation disabled. A second required test, S7-C2, ran the S7-P0 system on the stereo Sennheiser data to calibrate the baseline performance on the matched channel with most of the noise suppressed.

All data was run for each condition, but separate error rates are reported for each of four microphone-environment combinations.

### S8. Calibrated Noise Sources

This Spoke exposed the system to two calibrated noise sources through the stand-mounted directional microphone used in S6 and S7. Stereo Sennheiser data was also collected for calibration. The two noise sources were pre-recorded music and talk-radio, set at SNR levels roughly corresponding to 20, 10, and 0 dB, measured through the stand-mounted microphone.

The test data was from the 5K-word WSJ prompts. The test mode was static SI recognition. As in Spoke S6, use of the stereo microphone-adaptation data was permitted to adapt to the channel.

The primary test demonstrated a noise compensation algorithm and the required S8-C1 contrast ran the same system with the compensation disabled. A second required test, S8-C2, ran the S8-P0 system (compensation enabled) on the stereo Sennheiser data to calibrate the compensation algorithm on the matched channel with most of the noise suppressed. An additional optional test, S8-C3, ran on the Sennheiser data again but with the S8-C1 system (compensation disabled) to set the baseline.

All data was run for each noise source and SNR, but separate error rates are reported for each of six noise-level combinations.

The data sets were calibrated by adjusting the noise source levels with the subject in place until a sample set of utterances produced the specified SNR levels as measured by software supplied by NIST. But due to the variability of the sources themselves and the observed tendency of the subjects to drift in the level of their response over the session, the SNR levels measured on the test data differ considerably in places from the target SNR.

### S9. Spontaneous WSJ Dictation

The purpose of Spoke S9 was to simulate the application of large-vocabulary CSR to the dictation of news stories. The data was collected from practicing print-media journalists who composed stories spontaneously as described in [1]. News topic were chosen at the discretion of the subject after a priming review of WSJ newspapers current at the time.

The primary test allowed any acoustic or LM training. There were two required contrasts. S9-C1 tested the S9-P0 system on the H1 data, measuring the change in performance on read WSJ due to generalizing toward spontaneous data. The S9-C2 test ran the H1-C1 (controlled Hub baseline) on the S9 data to calibrate the difficulty of the spontaneous data.

## 4. Summary

In the first trial of the Hub and Spoke evaluation paradigm in November, 1993, 11 research sites participated, including 5 sites outside the ARPA community. The result was a rich array of comparative and contrastive results on several important problems in large vocabulary CSR, all calibrated to the current state-of-the-art performance levels. A complete listing of the numerical results can be found in [2]. For interpretive results, the interested reader should consult the comtemporary papers of the participating sites.

Two cautions are in order when attempting to interpret these results. First, since the acoustic training and development test data were distributed quite late, and since the Hub and Spoke paradigm was under development up to two months prior to the evaluation, a considerable burden was imposed on the participants who were rushed through the data processing and system training steps and were often denied a complete understanding of the rules. Some anomalies in the results did occur due to these undesirable circumstances.

Secondly, it's important to remember that the only tests for which fair and informative direct comparisons can be made across systems (and sites) are the controlled C1 contrasts for either of the two Hub tests. All other tests are designed to produce informative comparisons only within a given system run in two contrastive modes. So in general, only *intra*-system comparisons should be made on the Spoke tests.

The Hub and Spoke evaluation paradigm appears to have met the competing requirements of supporting the variety of important research interests within the ARPA CSR community while providing a mechanism to focus that work into well-defined and competitively charged evaluations of enabling technology. It is a flexible framework that encourages work in diverse problems in CSR.

It is also a very structured framework that treats all tests conducted in an evaluation as if they were scientific experiments, specifying controls where appropriate to maximize the amount of information contained in the results. This structure also helps keep the effort of the participating research community focused around a productive core of problems. If the Hub and Spoke paradigm is to be truly successful, however, it will need to sustain that focus over time in a manner analogous to the very successful Resource Management based evaluations of the late 1980's.

### Contact Information

The complete specification of the 1993 evaluation is contained in a documentation file included with the 1993 evaluation data distributed by NIST. It can also be obtained by sending a request to Francis Kubala via e-mail to fkubala@bbn.com.

Sites that are interested in participating in future ARPA-sponsored CSR evaluations can notify David Pallett at the National Institute of Standards and Technology (NIST). E-mail: dave@jaguar.ncsl.nist.gov.

The evaluation test data, as well as the training and development test data used in this evaluation, and all accompanying documentation, are available from the Linguistic Data Consortium (LDC). E-mail: ldc@unagi.cis.upenn.edu.

## Acknowledgement

For their dedication in producing the test data required to support the 1993 Hub and Spoke evaluation, the CCCC would like to thank Denise Danielson and Tom Kuhn at SRI as well as Jack Godfrey and Dave Graff at LDC. Without their committed work this ambitious evaluation would not have been possible.

## References

1. Bernstein, J., D. Danielson, "Spontaneous Speech Collection for the CSR Corpus", *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Feb. 1992, pp. 373-378.
2. Pallett, D., J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Pryzbocki, "1993 Benchmark Tests for the ARPA spoken Language Program", *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann Publishers, Mar. 1994, elsewhere these proceedings.
3. Paul, D., J. Baker, "The Design for the Wall Street Journal-based CSR Corpus", *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Feb. 1992, pp. 357-362.