Measurements IN Support OF Research Accomplishments

How the National Institute of Standards and Technology provides a basis for evaluating these novel systems.

David S. Pallett, John S. Garofolo, and Jonathan G. Fiscus

Since 1987, the National Institute of Standards and Technology (NIST) has provided support to other federal governmental agencies' speech research programs, and the university and industrial research communities, by developing, and implementing, test protocols, and reference test sets to provide quantitative measures of the progress of research accomplishments. These test sets are distributed to participants in the course of NIST's implementations of community-wide tests, and the results of the tests are widely quoted. Throughout this period, progress was traditionally measured in terms of word error rate—the percentage of substitution, deletion, and insertion word errors, relative to the number of words in an accurately transcribed set of recorded speech test material. Word error rates have shown a marked decline as the technology has improved and additional materials become available for system training. Program managers have generally challenged the research community to undertake more difficult tasks, as it becomes evident that substantial progress has occurred.

Figure 1 shows some of the historic word error rate data associated with the development of large vocabulary continuous speech recognition technology within the Defense Advanced Research Projects Agency (DARPA) research community. From approximately 1992, research results were reported on speech recorded from "read" texts derived from electronic newswire services (the *Wall Street Journal* and *North American Business News* corpora). These corpora provided a well-controlled reference condition, but limited the challenge to the research community because of the absence of other phenomena such as spontaneity, the use of different microphones, tele-

phone channel effects, competing background noises and music, and so forth.

It was recognized that radio and television broadcasts represented an alternative and rich source of potential real-world development and test materials. In 1995, a pilot study was conducted using materials derived from broadcasts of Public Radio International's "Marketplace" programs to gauge the state-of-the-art in the automatic transcription of this business news broadcast. The lowest reported word error rate (27%) was judged encouraging enough to initiate the Linguistic Data Consortium's efforts to acquire the intellectual property rights for a large corpus of radio and television broadcast news, subsequently termed the "Broadcast News Corpus." NIST subsequently developed and implemented benchmark tests using the broadcast news corpora in 1996, 1997, and, most recently, 1998. The lowest reported overall test set word error rate in the 1998 tests was 13.5%—half that reported for the 1995 "Marketplace" tests.

Information retrieval researchers began using these large collections of broadcast news materials as a resource for research. As a result, a new area of research focus developed, termed "spoken document retrieval," in view of the observation that news broadcasts largely consist of "stories," sometimes with texts adapted from concurrent newswire releases. Another outgrowth of the research community's interest in the availability of large quantities of broadcast news was the development of a Topic Detection and Tracking (TDT) research program. From the perspective of this research community, technology is to be developed to study the news media's coverage of topics, in which a topic is defined as a seminal event or activity, along with all directly related events and activities.

Figure 1. History of speech recognition benchmark tests implemented by NIST. (Lowest word error rate reported.)

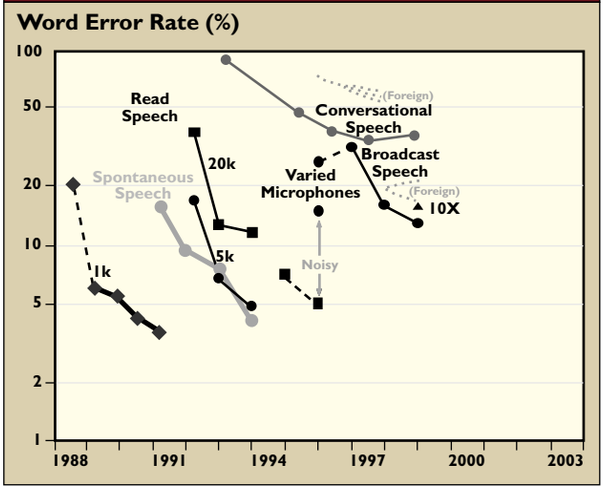
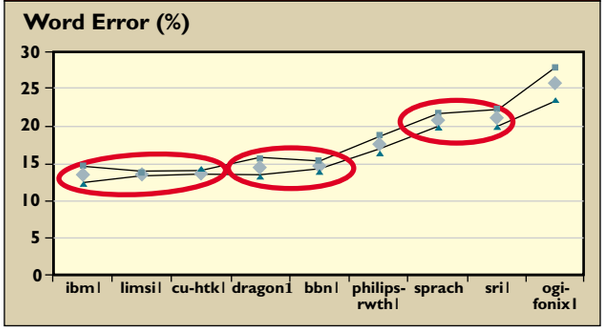


Figure 2. Results of the November 1998 DARPA Broadcast News Benchmark Tests (baseline systems). Ovals indicate that differences in word error rate, for the enclosed systems, are not significant, using the NIST MAPSSWE significance test.



1998 Broadcast News Tests

For the 1998 broadcast news transcription tasks, NIST provided participants with a test set consisting of two 1.5-hour subsets, each of which consisted of randomly selected story-length segments, drawn from materials provided to NIST by the Linguistic Data Consortium [4]. There were two tasks associated with the use of this test material: implement automatic speech recognition technology without the imposition of any computational constraints "in less than 10 times real time," and to implement automatic speech recognition technology with subsequent tagging of named entities—an information extraction task. This last task involved the participation of researchers from the Message Understanding Community (MUC) at MITRE and SAIC.

Results of the 1998 Broadcast News Tests

Figure 2 indicates the results of the unconstrained or baseline systems. Note that although an IBM-developed system yielded the lowest overall word error rate (13.5%), the application of statistical significance tests indicates that differences in performance between the IBM, the French National Laboratories' Laboratoire d'Informatique pour la Mechanique et les Sciences de l'Ingenieur (LIMSI) and Cambridge University's HTK (CU-HTK) systems were not significant. Thus the figure shows that excellent performance was achieved at several sites, both domestic and abroad.

For the "less than 10 times real time" systems, the lowest word error rate was 16.1%, achieved by a collaborative effort involving Cambridge University's

HTK group and Entropics, Ltd. This performance can be contrasted with the word error rate achieved by the best performing baseline system (13.5%)—a modest relative increase in error rate of 19%—and a remarkable reduction in runtime from approximately 2,000 times real time for the best performing baseline system to approximately 10 times real time—a reduction of 200-fold.

For the information extraction task, system developers were challenged to correctly recognize and tag the test set's named entities (locations, organizations, and persons), and also numeric quantities (money and percentages) and expressions involving times (date and time).

A complex testing protocol was developed. This involved the application of the BBN-developed IdentiFinder entity tagging software to not only the human-generated reference transcripts, but also to all of the hypothesis transcription files submitted for the baseline and less than 10 times real time systems. Our use of the IdentiFinder software on the hypothesis files provided a baseline to evaluate recognition focusing on entities that were of special significance in information extraction. The IdentiFinder software achieved F-measures of 82% for several of the well-performing ASR systems. This test protocol enabled investigation of the relationships between the information retrieval measure (the "F-Measure"—a weighted combination of Precision and Recall) and several different measures of word error rates, and the effect of error rate on tagger performance. We also implemented a contrastive test in which sites were free to implement both ASR and entity tagging. For the case where systems included both site-developed ASR systems and site-developed taggers, a BBN-developed system achieved an F-measure of (also) approximately 82%, with a word error rate of approximately 14.7% [5].

The availability of large collections of broadcast news and newswire data enables the study of the detection and tracking of news coverage of topics.

1998 Spoken Document Retrieval TREC Task

The technical challenge that the Spoken Document Retrieval (SDR) community addresses involves the "search and retrieval of relevant excerpts from collections of recorded speech," as an initial step toward information access in multimedia materials SDR involves the integration and development of two core technologies—speech recognition and information retrieval. The first SDR tests were implemented by NIST and reported at the Sixth Text REtrieval Conference (TREC). In 1998, the TREC-7 Conference included an expansion of the task from "known item" to "ad hoc-style" topics, and a doubling of the size of the corpus (from approximately 50 hours to approximately 100 hours) [2].

Figure 3 shows a block diagram including the major elements of the spoken document retrieval process. A corpus of broadcast news recordings provides data to a broadcast news speech recognition engine, which is tasked with generating transcripts. The transcripts are processed by an IR search engine, which treats the transcripts as a text collection (making use of manual story segmentation information), and generates indices for the stories in the transcripts. Topics (queries) are fed to the IR search engine, which produces a ranked document list, temporally linked back to the broadcast news corpus. Measures that are applied to this process include word error rates for the speech recognition subsystem, and mean average precision for the retrieval subsystem, with the use of human-generated relevance assessments for a set of queries.

The 1998 SDR tests, using an approximately 100-hour broadcast news corpus, involved a set of 23 ad-hoc style queries, or "test topics." Queries ranged from "Find reports of fatal air crashes?" to the more verbose "What economic developments have occurred in Hong Kong since its incorporation into the Chinese People's Republic?" The number of relevant stories in the corpus for each topic ranged from 1 to 60, with a mean number of approximately 17.

Several of the participating sites also provided the output of their ASR systems, so that these could be scored, and so that the other participating sites could investigate retrieval using these outputs in a cross-recognizer condition. The primary purpose of the test was to evaluate retrieval accuracy.

The traditional TREC information retrieval performance measure has been Mean Average Precision (MAP). The highest MAP of the SDR systems oper-

The test protocols and test data sets have proven to be of inestimable value in the course of the development of the core speech and language technologies.

ating on reference (perfect) transcriptions was 0.567, achieved by a group of researchers at the University of Massachusetts. The same group achieved MAP of 0.508, operating on texts produced by a Dragon Systems recognizer, while a group of AT&T researchers achieved MAPs of 0.507 and 0.512, operating on texts produced by their own recognizer.

A number of different procedures for scoring speech recognition error rates were investigated by NIST since the 1998 TREC meeting, in order to investigate the correlations between the measure used by IR researchers, MAP, and different ASR error metrics. Not surprisingly, Named Entity based word error rate measures were found to be the most highly correlated with Mean Average Precision. Surprisingly, for a wide range of word error rates, the MAP declines very gradually with increasing word error rate from the ASR systems—suggesting that even very imperfect ASR technology may be of substantial value in building SDR systems.

Figure 3. Spoken document retrieval process showing recognition and retrieval components.

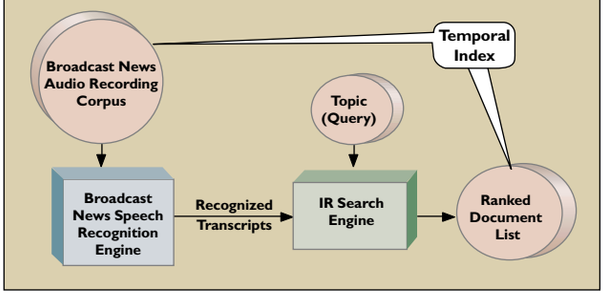
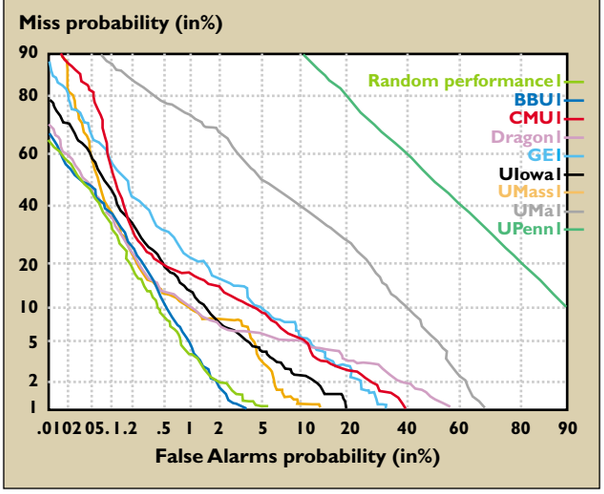


Figure 4. DET curve of 1998 TDT primary tracking systems (newswire and ASR text sources, four training stories, given boundaries for ASR).



1998 Topic Detection and Tracking Tests

The availability of large collections of broadcast news and newswire data enables the study of the detection and tracking of news coverage of topics. For our purposes here, a topic is defined as a seminal event or activity, along with all directly related events and activities. The LDC's TDT2 corpus comprises a set of annotated newswire and radio and television broadcasts, from six sources collected over a six-month period, including approximately 54,000 news stories. Participants in recent TDT tests implemented by NIST [1] were challenged to perform any of three well-defined tasks:

- Find the story boundaries, for radio and TV materials—the segmentation task;
- Find all the stories that discuss a given target topic—the tracking task; and
- Associate together all of the stories that discuss a topic, for all topics—the detection task.

Figure 4 indicates one of the analyses involved in the TDT program. The figure represents a Decision Error Tradeoff (DET) plot, showing, for the several systems, the tradeoff between the miss probabilities vs. the false alarm probabilities for a topic tracking experiment involving the use of both newswire and ASR-generated texts for four training stories and given boundaries. Better performance, in general, is shown by results that approach the lower left corner of the normal deviate scaled figure. In this case, the system developed at the University of Pennsylvania had the best performance.

Development of News on Demand Systems

Several of the sites from which developers of News on Demand (NOD) systems described in this issue

are participants in the broadcast news tests described here (BBN, CMU, MITRE, and SRI). Participation in these tests has provided a stimulating environment for productive technical discussions. The NIST-developed test protocols and test data sets have proven to be of inestimable value in the course of the development of the core speech and language technologies.

However, the breadth of scientific disciplines encountered in building a NOD system is great. As one researcher [3] noted "These demonstration systems integrate various diverse speech and language technologies including (not only) ASR (but also) speaker change detection, speaker ID, name extraction, topic classification, and information retrieval. Advanced image processing capabilities are also demonstrated on the news video including scene change detection, face ID, and optical character recognition from the video image."

The challenge to developers of NOD systems, and that of the larger research community, is to incorporate the scientific knowledge gained in these broadcast news corpus-based speech and language tests into their demonstration systems. C

References

1. Doddington, G., et al. Topic detection and tracking: TDT2 overview and evaluation results. In *Proceedings of DARPA Broadcast News Workshop*, (Feb.–Mar. 1999).
2. Garofolo, J., et al. 1998 TREC-7 spoken document retrieval track. In *Proceedings of DARPA Broadcast News Workshop*, (Feb.–Mar. 1999).
3. Kubala, F. Broadcast news is good news. In *Proceedings of DARPA Broadcast News Workshop*, (Feb.–Mar. 1999).
4. Pallett, D.S., et al. 1998 broadcast news benchmark test results: English and Non-English word error rate performance measures. In *Proceedings of DARPA Broadcast News Workshop*, (Feb.–Mar. 1999).
5. Przybocki, M., et al. 1998 broadcast news evaluation information extraction named entities. In *Proceedings of DARPA Broadcast News Workshop*, (Feb.–Mar. 1999).

David S. Pallett (dpallet@nist.gov) is the manager of the Spoken Natural Language Processing Group at the Information Technology Laboratory at the National Institute of Standards and Technology in Gaithersburg, MD.

John S. Garofolo (jgarofolo@nist.gov) is a computer scientist with the Information Technology Laboratory at the National Institute of Standards and Technology in Gaithersburg, MD.

Jonathan G. Fiscus (jfiscus@nist.gov) is a computer scientist with the Information Technology Laboratory at the National Institute of Standards and Technology in Gaithersburg, MD.