

# A LOOK AT NIST'S BENCHMARK ASR TESTS: PAST, PRESENT, AND FUTURE

David S. Pallett

National Institute of Standards and Technology (NIST)  
Gaithersburg, MD 20899

## ABSTRACT

*This paper discusses the role that the NIST Speech Group has played by coordinating and implementing objective benchmark tests in the automatic speech recognition research community. From 1987, when the first tests were implemented, to the present time, at which the Rich Transcription concept has been developed, the tests have served to document the state-of-the art. Testing has involved a number of different, but complementary, domains. These test results document the progress of the technology – ever-lower WERs with continued attention to the task domain of the era, and several changes of focus to address ever-more challenging tasks,*

## 1. PREAMBLE

In 1981, the author's attention was brought to an "evaluation of equipment" by George Doddington and Thomas B. Schalk, at Texas Instruments [1]. The tests described in that article dealt with tests conducted on speaker-dependent recognizers from seven different manufacturers on a 20-word isolated word vocabulary involving 16 speakers. The speech database, subsequently known as the "TI 20 Word" database, was developed at TI for the purpose of providing a basis for objective testing. A speaker-independent HMM-based system developed by Verbex had the lowest overall Word Error Rate (WER). These results were discussed at an informal session at the 1981 ICASSP Conference, and attracted great interest. Subsequently, a "Group on Speech I/O Systems Performance Assessment" was formed, which eventually led to the publication of an article in the Journal of Research of the National Bureau of Standards [2].

Subsequently, Doddington made the TI speech data available to the research community through the National Bureau of Standards.

NIST's involvement with DARPA speech recognition programs started in 1984, with interest shown in providing

objective quantitative measures of performance. The first publication describing NIST's benchmark tests appears in the Proceedings of the Speech Recognition Workshop sponsored by DARPA in February 1986 [3]. A Strategic Computing draft document was developed at DARPA in December, 1985 (using a draft of the NIST article) that identified key issues in some detail. Representatives of BBN, CMU, Dragon System, MIT and TI participated in discussions that developed the test protocols.

In the early phases of DARPA-sponsored research, two key speech databases were collected at TI. One of them, known as the TIMIT Acoustic-phonetic database, involved a collaboration between TI and the staff of the MIT group led by Victor Zue. It established "standards" for speech data – 16 kHz sampling rate with 16 bit quantization, and the use of a close-talking, headset-mounted Sennheiser microphone [4]. A second speech database that TI collected in the early phases of the research has become known as the "Resource Management" speech database. This corpus was developed in order to support a focused speech recognition research program and to support objective evaluations of ASR systems.

## 2. PAST

### 2.1. The Resource Management Era

The creation and availability of the Resource Management corpus lead the way for NIST benchmark tests. It was based on language believed to be useful in managing naval resources using spoken language. It was limited to a 991 word lexicon and using a "pattern grammar" with only approximately 2800 sentence patterns defined. The Resource Management speech database texts were read by speakers at TI, using natural continuous speech. Researchers were challenged to develop systems that processed the speech data, and generate textual output. NIST had the role of defining the Standard Normalized Orthographic Representation (SNOR) that was used until recently, conventions such as the omission of punctuation,

breath noises, etc, and developed scoring software from prototypes provided by TI and BBN

The first Resource Management tests, conducted in early 1987, involved the BBN BYBLOS HMM-based system, and, at CMU, a system involving signal processing, acoustic-phonetic, lexical access and parsing modules.

The benchmark tests implemented by NIST are open to all research sites. Test participants in the NIST benchmark

tests must submit an informative system description document along with their test data, and make presentations about their approach at Workshops that are organized by NIST. These tests have involved many sites from many nations.

The results of the tests that NIST has performed over the years are shown in Figure 1. In all of these data, the performance of the best-performing system at that time is shown.

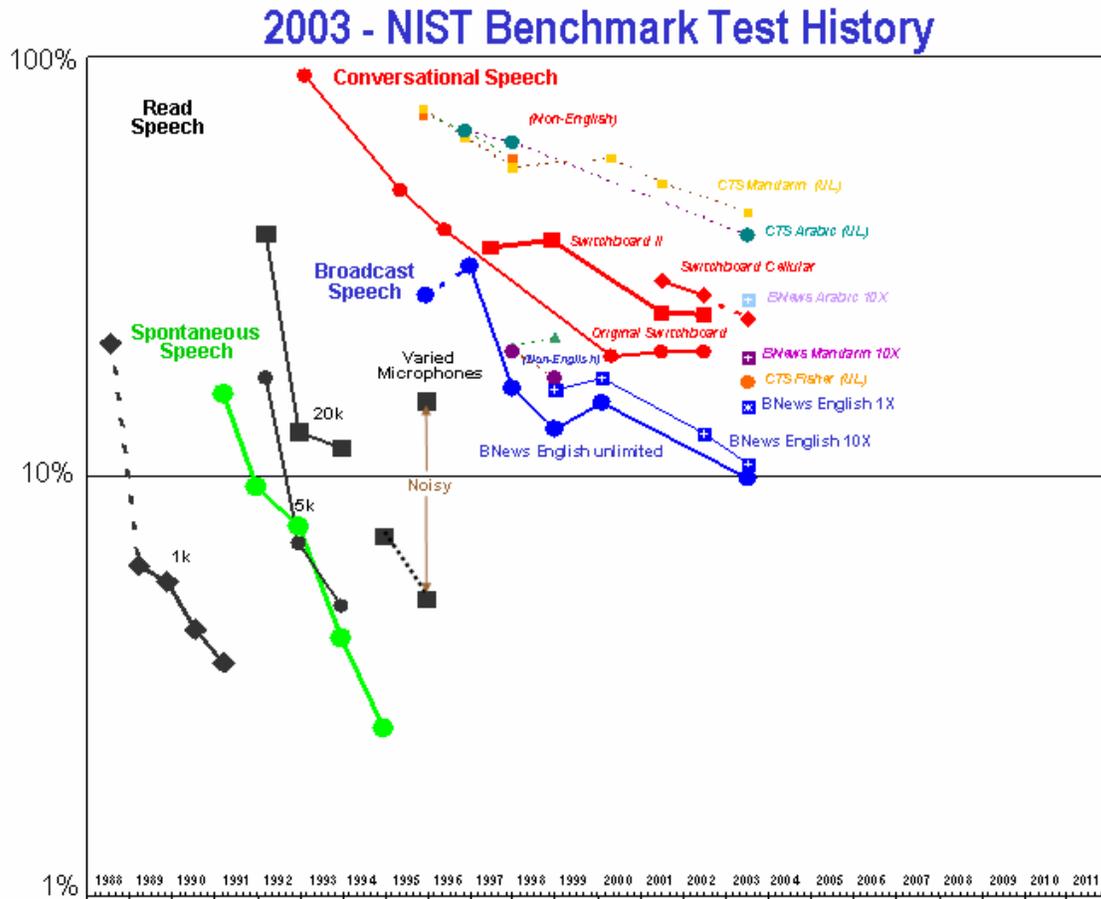


Figure 1 NIST Benchmark Test History

The first data point, for June 1988, with a WER of 20.7%, represents the state-of-the-art at that time using the Resource Management database. Over the several years of research with this task, WERs decreased to 3.6% in February 1991. This pattern, of reductions in WER with continued research and development on a focused task, has been repeated over and over throughout the years.

Not shown in this figure are results for the final tests conducted in another DARPA program, involving a test set that was not as carefully selected (with regard to the

distribution of gender and dialect history) [5]. WERs for these tests turned out to be higher than in the previous Resource Management tests. This raised our awareness to the variabilities between test sets.

## 2.2. The Air Travel Information Systems (ATIS) Era

Another challenge that was put to the research community early on in the DARPA program (contemporaneously with the Wall Street Journal) was to develop technology providing access to information – in this case, information

about air travel - using spontaneous speech. The challenge was to “understand” spoken queries in a human-machine dialogue, not to simply generate a transcription. Thus the ATIS task became a challenge to both the speech recognition and synthesis communities, but also to the natural language processing communities. This became known at the ATIS task.

A “frozen” relational database was developed that provided information on available flights, and users were asked to request information about air travel. A typical query might have been “I’d like some information about flights between DALLAS and BOSTON”. In this representation, the information that was key to the query is shown in capital letters, and the “carrier phrase” is in lower case letters. The realization that the number of cities was limited, and that simple language modeling could be used, greatly simplified the challenge. Systems that were developed typically had vocabularies of several thousand words.

NIST, in conjunction with the researchers, developed a “Principles of Interpretation” document that defined the correctness of responses, with both minimally correct answers, and a penalty for verbosity.

A portion of the ATIS speech data was collected at the several sites participating in the research, with the efforts of a Multi-site Atis Data Collection Working (MADCOW) group. This data was collected, transcribed and scored at the sites, and distributed to the community by NIST. The ATIS data was obtained with the use of a “wizard of oz” paradigm, and the data was recorded using the “standard” headset-mounted microphone.

WERs for this spontaneous speech ranged from 15.7% in February 1991 to 2.5% in November 1994, again marking the improvement in the technology over time.

### 2.3. The Wall Street Journal Era

The first use of large textual databases for statistical language modeling purposes made use of textual data obtained from the Wall Street Journal (WSJ). Text files derived from the WSJ were selected by TI staff with consideration of the vocabulary size required for coverage – 5000 and 20,000 word lexicon sets. These texts were read by speakers at TI and used for research and testing.

WERs of 17.1% for the “5K” lexicon material, and 32.9% for the “20K” materials were reported in February 1992,

The scope of the texts available for statistical language modeling was expanded to include other “North American

Business” (NAB) news sources, and the task then became referred to as the WSJ/NAB news task.

In 1995, data was collected at NIST with an assortment of microphones situated in the vicinity of the terminal on which the texts to be read were displayed. These were used to test the sensitivity of the acoustic modeling of the systems, and while a WER of 5.1% was achieved with the close-talking reference microphone that had been used in previous tests, the WER rose to 15.1% using a varied assortment of alternative microphones.

Note that in the WS/NAB era the previously noted trend toward reduced WER again occurred.

### 2.4. The Broadcast News Challenge

It was recognized by NIST that the “Marketplace” broadcasts produced and distributed by Public Radio International dealt with much the same subject matter as the WSJ/NAB texts. One broadcast was recorded and processed at NIST using a version of the BBN BYBLOS recognizer, and the results were scored and discussed with DARPA management. A decision was made to address the “Broadcast News” challenge, and arrangements were made by the LDC to obtain rights to record, transcribe, and distribute broadcast news data to the research community.

In the early phases of this challenge, it was recognized that the data exhibited considerable diversity – ranging from the fluent, apparently read, speech of news anchors, to spontaneous, disfluent, speech collected in various potentially noisy environments. Thus a decision was made to report scores in several “focus conditions”, corresponding to the diversity that had been noted, in addition to overall test set scores. Another decision was made to disregard the challenge of transcribing sports material and commercials

It has been noted that it is difficult to ensure that each test set presents the same challenge to the system developers for a particular clearly defined transcription task. Test sets vary in many dimensions. The relative amount of the “easy” fluent speech, typically from news anchors and apparently read, varies from one broadcast to another, and from program source to program source, and as well there is variation in the degree of spontaneity and the amount of the background noise. We can see this effect when comparing the first “dry run” broadcast news WER, 27%, in November 1995, using only Marketplace data, to the results in the tests of 1996, a WER of 31.8%, with essentially the same systems. The 1996 tests involved four

sources, two radio broadcasts and two television broadcasts, not just the one source.

Cross-site comparisons are certainly valid using any given test set. Comparisons over several years are somewhat less valid as indicators of progress in the development of ASR technology, because of the variability in test set properties that can't be controlled with real, human-human data. Using appreciably larger test sets, to improve statistical sampling considerations, might reduce the variability, but these large test sets would probably be unacceptable to the research community. These considerations later led to a decision to reuse test material in order to track progress.

The lowest "broadcast news era" WER (in the 1998 tests) shows a WER is 13%.

The Broadcast News era started with WERs of 31% in 1996, experienced a low WER of 13% in 1998, and ending the era, in 1999 with a WER of 15%. Once again, with continued focus on the task by the research community, the trend toward improved technology has been shown.

## 2.5. The Communicator Era

Another DARPA program that NIST was involved in was the "Communicator" program. (WER results for this program are not shown on Fig. 1). This program was in some sense a continuation of the ATIS task spoken language understanding task, but involved "real" online, not "canned" or "frozen" air travel information sources, the interface with the technology was via the telephone and "wizards" were not employed. In addition to air travel information, system developers were challenged to provide information about hotel and rental car availability for some of the queries. Transcription and scoring of system responses were performed by the participating sites.

NIST's role in this research program included providing a subject pool, and in collecting and distributing data from the several sites. In one year's activities, NIST maintained a data collection system that monitored and recorded the mixed-initiative dialogue transactions, and in a second year, NIST collated data from the participants and performed some analyses of the data.

Although the primary focus of the Communicator program was not solely to develop improved telephone-based speech recognition systems, the median WER was reduced to 17% in the second year of the program for this task

NIST investigated the relationship between WER and several measures of task success and user satisfaction. We found that, as expected, measures of task completion improved as WER decreased, and that when speech recognition deteriorates, so do user satisfaction and efficiency. However, overall performance differs across systems even when speech recognition accuracy is equal. Thus speech recognition accuracy alone is not the whole story in a spoken dialogue system in which users interact via speech alone to perform a task.

## 2.6. The Conversational Speech Challenge

In an early endeavor, TI collected what has become known as the "Switchboard" spontaneous speech database. This database consisted of recordings of participants in a telephone-based discussion of "topics" suggested by an automaton that served to connect the participants and record the conversations. The original purpose of the database was to support research in topic and speaker spotting. This collection paradigm was repeated at other sites in other years, frequently including discussions between a college-student age subject population. More information on the Switchboard data can be obtained from the Linguistic Data Consortium [6].

First use of the Switchboard data for speech-to-text purposes was in January 1993, with a reported WER of 90%. By April 1995, WER had improved to 48%, and WER had, by 2001, apparently plateaued to the vicinity of 19%, for original Switchboard data. However, later phases of the Switchboard data collection had included the collection of data using cellular phones, and WERs for cellular data are shown to be higher, 29.2% in May 2001 and 27% in 2002.

Conversational speech has also been collected by the LDC in languages other than English, and used for tests as shown in the dotted-line data. Reporting WERs in foreign languages is somewhat complicated by the necessity of defining transcription and scoring conventions for languages that do not use conventional Western orthography and for which the concept of a "word" is defined differently. For example, we report "character error rates" for Mandarin.

Attention should be paid to the processing times associated with any set of results. In the earlier tests discussed in this paper, processing times were essentially unlimited, and the organization with the most processing power, or time to accommodate the tests, had a distinct advantage. In 1998, attention was, for the first time, focused on reducing, and documenting the processing times allowed for the tests. A target for a "faster" system

was 10 times real time (10X). WERs that were higher than for the unlimited time systems are typically encountered, and for real time (1X) systems even higher error rates prevail.

### 3. PRESENT

#### 3.1. The Rich Transcription Era

In 2003, NIST implemented the first tests in the DARPA Effective, Affordable, and Reusable Speech-to-text (EARS) Program.

The goal of this DARPA Program is to provide a “Rich Transcription” – providing not just a text stream, but a rich transcript that includes metadata - as the output of an ASR system. Therefore the terminology changed somewhat – the overall challenge of Rich Transcription includes both Speech To Text (STT) and Metadata Extraction (MDE). In the STT portion, there is a focus on the core ASR technology.

The initial participants in the EARS Program that are to focus on the core technology include one team lead by BBN (but also involving LIMSI), Cambridge University, IBM, and another team involving SRI, ICSI, and the University of Washington. Another team that is to focus on metadata extraction involves the MIT Lincoln Laboratory group. There is also a “Novel Approaches” multi-site team led by ICSI. Linguistic Data is to be provided by the LDC, and NIST has a role in the evaluation infrastructure.

The accompanying metadata is useful in increasing the readability of the transcripts. Initial focus of the metadata extraction (MDE) effort includes disfluency recognition (marking verbal fillers such as filled pauses, discourse markers, and verbal edits), semantic unit segmentation, and speaker diarization (marking the times corresponding to speaker changes, and providing speaker identification information). The focus of the EARS efforts are on both Broadcast News.(shown in the figure as (BNews)) and Conversational Telephone-based Speech (CTS), in English and Chinese and/or Arabic.

NIST is implementing tests using two test set components: (1) a “current” era component, drawn from current broadcast news and recently collected conversational speech, and (2) what has been termed a “progress” test set that is to be re-used over a period of several years. Because of its re-use, the reference transcriptions used for scoring test submissions are not to be released by NIST.. Principal investigators at participating sites have been required to assert in writing that the test data, and all

derivative files have been removed from that site’s systems when the results of processing the test data have been submitted to NIST. Use of this progress test set will enable better tracking of progress in the development of Speech To Text (STT) algorithms, while the use of the current era test set should continue to provide valuable inter-site comparisons, and permit tracking the changing focus of current news.

A considerable amount of new infrastructure activity has been involved in preparing for the EARS Program. Not only did the community need to agree on a format convention and scoring software, but it has been necessary to reach agreement on issues such as the time intervals associated with scoring speaker segmentation, whether or not to address overlapping speech, etc.

The LDC’s data collection paradigm for CTS has changed. Data are currently being collected using what has been termed the “Fisher” system, which is an adaptation of the Switchboard technology. When in operation, this system constantly tries to reach participants who have enrolled over the telephone and then, once an individual speaker has been contacted, the system attempts to connect with another participant, and initiate recording of the CTS.

The first benchmark tests of the EARS era were conducted in early 2003. A second phase of the testing – focusing on metadata extraction – is to take place in the Fall of 2003.

The STT portion of the tests that were conducted in early 2003 are shown as the right-most data points on Figure 1. In the most recent (2003) tests:

The highest error rate (in this case, character), 42.7%, was for Mandarin CTS.

Next highest (37.5%) was for Arabic CTS.

Next highest (26.3%) was for BNews in Arabic, using a “10X” system.

Next highest (23.8%) was for Switchboard cellular CTS.

Next highest (19.1%) was for a BNews Mandarin 10X system.

Next highest (16.7%) was for recently collected CTS (using the LDC’s “Fisher” paradigm) data.

Next highest (14.6%) was for a BNews English language 1X (i.e., real-time) system.

Next highest (10.7%) was for a BNews English language 10X system.

The lowest WER (9.9%) was obtained with an English language BNews system with unlimited computational power.

In summary, it may be worthwhile to note that the performance on English language data is better than on the foreign languages, that it is better on BNews than on CTS, and that faster systems, in general, perform worse than slower systems.

## 4. FUTURE

### 4.1. The Meeting Room Era

One of the next big challenges in automatic speech recognition (ASR) is the transcription of speech in meetings. This task is particularly problematic for current recognition technologies, because in realistic meeting scenarios, the vocabularies are not constrained, the speech is spontaneous and overlapping, and the microphones will be inconspicuously placed.

To support the development of meeting recognition technologies by the ASR research community, the NIST Speech Group is providing a development and evaluation infrastructure including: richer transcription and annotation conventions, a corpus of audio and video from meetings collected at NIST using a variety of microphones and video cameras, new evaluation protocols, metrics, and software, sponsoring workshops, facilitating multi-site data pooling, and helping bring the community together to focus on the technical challenges.

We feel that this data will provide an abundance of challenges to the ASR research community. The simplest challenge to be addressed would be to generate a traditional transcript – without necessarily providing metadata – using the head-mounted microphone data. Alternatively, the signals from the head-mounted microphones could be combined to create a relatively easy multi-participant challenge that would include overlapping speech. Even more challenging would involve use of the table-mounted microphones. Adding the Rich Transcription’s metadata into the picture might be another interesting challenge. This progression of difficulty would challenge acoustic modeling in ways that current research doesn’t.

We hope that sponsorship of research directed to this challenge will ensue.

## 5. ACKNOWLEDGEMENTS

The author would like to acknowledge the contributions made by George Doddington and Janet Baker, who first suggested the role of benchmark tests that NIST might administer, and John Makhoul and Raj Reddy, whose organizations were the first to participate in these tests. The cooperation of the researchers at the sites participating in our tests is gratefully acknowledged. The tests could not have had the role they have had without the sponsorship of several DARPA Program Managers and the NSA, and, of course NIST.

Finally, the staff of the NIST Speech Group deserve virtually all of the credit for administering the tests, and reporting on the results at many Workshops. Special credit is due to John Garofolo and Jonathan Fiscus, who developed scoring software and coordinated interactions with the many participants.

Further information about the NIST Speech Group can be found at <http://www.nist.gov/speech/>, and information about the speech databases cited in this paper can be found at the Linguistic Data Consortium’s web site: <http://www ldc.upenn.edu/>.

## 11. REFERENCES.

- [1] G. R. Doddington and T. B. Schalk, “Speech recognition: turning theory into practice”, IEEE Spectrum, Vol. 18, #9, pp.26-32, Sept. 1981.
- [2] D. S. Pallett, “Performance Assessment of Automatic Speech Recognizers”, Journal of Research of the National Bureau of Standards, Vol. 90, #5, Sept. – Oct. 1985.
- [3] D.S. Pallett, “A Benchmark for Speaker-Dependent Recognition using the Texas Instruments 20 Word and Alpha-set Speech Database”, in Proceedings, Speech Recognition Workshop, February 1986.
- [5] D.S. Pallett, J.G. Fiscus, and J.S. Garofolo, “Resource Management Corpus: September 1992 Test Set Benchmark Test Results”, in Proceedings of the ARPA Microelectronics Office Continuous Speech Recognition Workshop, Stanford CA, Sept. 21-22, 1992, pp. 1–18.
- [6] See, for example, <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=L-DC93S7-T>