USE OF BROADCAST NEWS MATERIALS
FOR SPEECH RECOGNITION BENCHMARK TESTS

David S. Pallett, Jonathan G. Fiscus, William M. Fisher, and John S. Garofolo
Spoken Natural Language Group, Information Technology Laboratory
Room A 216 Technology Building
National Institute of Standards and Technology (NIST)
Gaithersburg, MD 20899
E-mail: david.pallett@nist.gov

ABSTRACT

This paper reports on the use of materials derived from radio and television news broadcasts for research and testing purposes for large vocabulary Continuous Speech Recognition (CSR) technology. Tests using these materials have been implemented by NIST on behalf of the DARPA-funded speech recognition research community in 1995 and 1996, and are expected to continue for the next several years. Four research groups participated in the 1995 tests, and nine groups (at eight sites) participated in the 1996 tests. This paper documents properties of the training and test materials, describes a detailed annotation and transcription protocol that has been used for more than 100 hours of recorded data that has been made available through the Linguistic Data Consortium (LDC), and discusses test protocols and results of both the 1995 and 1996 Benchmark Tests.

1. INTRODUCTION

The use of speech data derived from radio and television news broadcasts was initiated in 1995 and 1996 in large-scale “benchmark tests” coordinated by NIST for DARPA. These data constitute what has been termed “found speech” -- speech recorded off-the-air -- not specially collected for research and tests. In November 1995, “dry run” tests were conducted by four sites using materials derived from radio broadcasts of a public radio series focusing on business news. In the tests implemented in November 1996, the scope of these tests was expanded to include both television and radio news broadcasts, and approximately 50 hours of transcribed materials were made available in the summer of 1996 by the Linguistic Data Consortium.

The task of successfully transcribing “found speech” as found in broadcast materials is of potential value not only as an interesting technical challenge for applications such as producing closed caption materials, but it has been recognized that even imperfect transcriptions will be valuable in applications involving audio indexing for multimedia documents and related document retrieval applications.

The 1995 “dry run” tests as well as the 1996 preliminary tests indicate the existence of a number of complicating factors in these tasks -- differences between “prepared” or “read” speech and “spontaneous” speech, transmission channel or bandwidth effects, the presence or absence of background sound conditions (music, voices, or noise), foreign-accented speech and combinations of these effects. For the 1996 tests, a detailed annotation convention was implemented by the LDC to capture these effects. This convention permitted the community to implement a particular form of partitioned evaluation, and for which NIST’s analyses of results could be partitioned into corresponding categories.

For all systems, perhaps most striking feature of these tests is that from segment to segment the word error rates sometimes vary over a wide range. The challenge that is presented by the broadcast materials is significant. In the 1996 tests, involving both radio and TV broadcast news materials, the system with the lowest measured error rate had an overall word error rate of 27.1%.

Plans are being made to continue this test series, and the development of improved technologies, for the next several years. Additional tests are to take place in November, 1997. The test procedures developed for these tests, and the supporting training and test data, will comprise de facto standards.

2. ANNOTATION CONVENTIONS

The procedures and conventions used for annotation of the broadcast news materials at the LDC are described by Graff in [2]. In addition to carrying out a transcription task, the transcribers were also assigned the tasks of: (1) marking the beginning and ending of each topical unit and identifying its type; (2) marking the beginning and ending of each speaking turn within a topical unit, identifying each speaker uniquely, and indicating (in the judgment of the transcriber) whether he/she is a native speaker of North American English; (3) for each speaking turn, indicating the channel quality and speaking mode (planned or spontaneous); (4) marking the beginning and ending points of three types of background sound conditions and subjectively characterizing the severity of these conditions; and (5) marking the occurrence of overlaps of adjacent speaking turns. The complex nature of these transcriptions was put into an SGML framework that is described in [2].

Experience within the community with the training materials indicated that special care was in order when transcribing and annotating the Evaluation Test Set. NIST staff collaborated with LDC staff in developing the reference annotated transcriptions for the test set. Garofolo, et al. [3] describe these procedures, which involved three transcribers annotating the test materials, followed by review and reconciliation of differences in the annotations by an "annotation reconciliation tsar". Similarly, in a subsequent pass through the data, three transcribers were provided with the reconciled annotations as the framework for transcription. The transcriptions developed by the three transcribers were then reviewed for differences and the differences "reconciled" by a "transcription reconciliation tsar" to produce the final annotated reference transcriptions used for the benchmark tests.

3. "UNPARTITIONED" EVALUATION (UE)

The test protocol implemented in the 1995 "dry run" tests required each system to process several unsegmented files corresponding to 15 to 30 minute broadcast segments. Each of the participating sites developed "segmentation" or "chunking" modules so that their systems could accommodate these unusually large speech data files. It became evident that the task of building segmentation modules presented its own challenges, possibly diverting research effort away from the speech recognition task.

Subsequent to the Workshop early in 1996, at which the 1995 results were discussed, several sites suggested the development of a "Partitioned" test protocol. Such a protocol would use partitioning information in accompanying annotations, in both processing the test data, and in summarizing the test results, to obviate the need for development of segmentation modules. Tests of complete systems that made no use of such information in the annotation files were to be regarded as "Unpartitioned" Evaluations, in contrast to "Partitioned" Evaluations for which test materials with similar properties were aggregated into subsets.

In the 1996 tests, three sites that had participated in the 1995 tests provided results for both Partitioned Evaluations (PE) and Unpartitioned Evaluations (UE).

4. "PARTITIONED " EVALUATION (PE)

For the 1996 tests, a number of focus conditions were defined for use in the PE and for NIST's analysis of the results. Particular combinations of conditions based on information in the annotations were defined. Table 1 [3] documents these combinations of dialect, speaking mode, fidelity and background conditions.

A substantial fraction of the material with native speakers of North American English falls into three of the focus conditions: (1) F0, the "Baseline Broadcast" condition characterized by planned speech with high fidelity (e.g., broadcast studio quality) and no background noise, and (2) F1, the "spontaneous speech" condition, also recorded in high-fidelity, low-noise conditions, but for which there is evidence of spontaneity and the presence of disfluencies can be noted, and (3) F2, the "Reduced Bandwidth" condition, including both planned and spontaneous speech, and possibly originating with telephone handsets and/or having been transmitted over telephone channels. As will be seen, typical performance of CSR technology is best for the F0 baseline condition, and increasingly less robust for F1 and F2.

Condition	Dialect	Mode	Fidelity	Background
Baseline Broadcast (F0)	native	Planned	High	Clean
Spontaneous Speech (F1)	native	Spontaneous	High	Clean
Reduced Bandwidth (F2)	native	(any mode)	Med/Low	Clean
Background Music (F3)	native	(any mode)	high	Music
Degraded Acoustics (F4)	native	(any mode)	High	Speech/ Other Noise
Nonnative Speakers (F5)	non-native	Planned	High	Clean
All other comb. (FX)	-	-	-	-

Table 1: 1996 Focus Conditions

5. TEST SET PROPERTIES

NIST implemented a detailed screening process in selecting the test materials and in preparing the reference transcriptions. The test set included two half-hour excerpts of radio news broadcasts (from PRI's "Marketplace" and NPR's "The World" broadcasts) and two half-hour excerpts from TV news broadcasts (from CNN's "Morning News" and CSPAN's "Washington Journal" broadcasts).

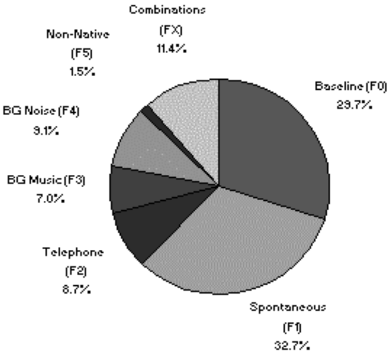

Figure 1: Test Material Distribution

Figure 1 [3] indicates the distribution of material across the several focus conditions. Note that only 1.5% of the test material falls into the F5 "Non-Native" speaker focus condition, possibly either an artifact of the small sample size, or an attribute of broadcast news in the United States.

6. ORTHOGRAPHIC TRANSFORMATIONS

A set of global mapping rules was implemented and used by NIST when processing test results in order to deal with lexical representations that were not to be regarded as errors. The rules covered four classes: contractions, alternate (or inconsistent) spellings, incorrect spellings that occurred in the training transcriptions, and compound words. As noted in [3] "The task of writing simple literal rewriting rules to expand apparent contractions with minimal over-generation turned out to be very difficult..." Alternate standard spellings were dealt with by referring to the American Heritage Dictionary, and in some cases, to Web searches to find commonly used representations for alternate spelling of people's names. Semi-automated tools (e.g., the use of spell-checkers using statistical language models) were developed and used to identify errors in the training transcriptions and to develop rules that forgive errors involving the mistranscribed words. In scoring, the final global mapping rule set included 348 transformations.

7. 1995 DRY RUN RESULTS: LESSONS LEARNED

Perhaps the most striking lesson learned from the 1995 tests is that the word error rates are exceptionally variable, frequently ranging from lows on the order of 10% to values in excess of 100%. This variability is attributed to the rapidly varying nature of the data as different speakers take turns and conditions change, and certainly due to the limitations on the robustness of current CSR technology. In the 1995 tests, the system with the lowest reported word error rates had an overall test set word error rate of 27.0%.

Figure 2 indicates the word error rates for individual partitioned segments throughout the course of a half-hour broadcast excerpt. Note that, in this broadcast, the lowest error rate, of ~3%, was noted with the speaker named (John) Dimsdale at ~270 seconds, just prior to a partitioned segment involving (President) Clinton, with a word error rate of ~26.3% for Clinton.

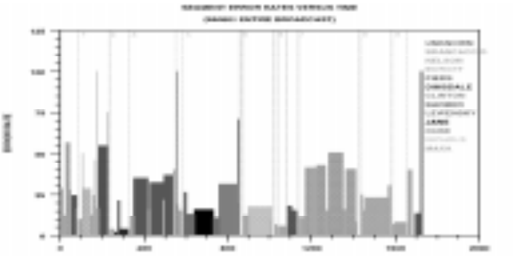

Figure 2: 1995 Word Error Rates

Analysis of the 1995 tests indicated the importance of differences in speaking style, transmission/microphone/handset bandwidth or "fidelity", and background noise in affecting error rates. These observations led to the annotation system used for the 1996 tests and to the development of the "Partitioned Evaluation" test paradigm.

8. 1996 TEST RESULTS

In the 1996 Broadcast News Benchmark Tests, nine different research groups, at eight sites, participated in the tests. In Table 2, the systems associated with these sites have the following designations: BBN Systems and Technologies (bbn1), Carnegie Mellon University (cmu1), the "Connectionist" and "HTK" Groups at Cambridge University's Engineering Department (cu-con1 and cu-htk1, respectively), IBM T.J. Watson Laboratories (ibm1), LIMSI/CNRS (limsi1), a collaborative effort involving New York University and SRI International (nyu1), Rutgers University (ru1 and ru2), and SRI International (sri1). Of these, BBN, CMU and IBM had participated in the 1995 "dry run" tests. Technical papers describing the research efforts involved in building these systems can be found in the *Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997.* See [2] for information about availability of this Proceedings.

DARPA CSR 1996 Broadcast News Hub-4 Benchmark Test

	Complete Test	F0	F1	F2	F3	F4	F5	FX
bbn1	30.2	21.6	29.5	32.7	23.3	38.4	31.8	49.9
cmu1	34.9	25.8	32.1	38.6	36.6	43.7	36.5	55.8
cu-con1	34.7	25.8	33.5	40.4	33.4	39.3	40.5	53.1
cu-htk1	27.5	18.7	26.5	33.1	23.6	29.1	21.7	51.0
ibm1	32.2	21.6	30.4	38.9	28.0	42.2	30.8	54.2
limsi1	27.1	20.8	26.0	27.1	20.3	33.3	27.8	46.1
nyu1	33.0	26.0	32.5	32.6	34.2	38.4	31.1	48.1
ru1	56.1	43.0	51.7	74.6	50.0	81.6	54.8	72.1
ru2	53.8	42.7	51.9	72.9	50.0	59.2	54.8	71.9
sri1	33.3	26.4	33.0	31.7	34.7	38.5	34.4	48.3

F0: Baseline Broadcast Speech
F1: Spontaneous Broadcast Speech
F2: Speech over Telephone Channels
F3: Speech in the Presence of Background Music
F4: Speech under Degraded Acoustic Conditions
F5: Speech from Non-Native Speakers
FX: All Other Speech

Table 2: 1996 Test Results

Table 2 [4] documents the complete test set word error rates as well as those for each of the "focus condition" subsets (F0 through FX). For the system with the lowest measured word error rate (limsi1) the word error rate for the complete test set was 27.1%, with error rates for the focus conditions ranging from 20.3% to 46.1%. Note that closely comparable results are reported for the cu-htk1 system. In [3] the results of NIST's implementation of paired-comparison statistical significance tests indicate that performance differences between the limsi1 and cu-htk1 are in effect not significant. As will be noted from the data of Table 2, in many cases, differences in performance between many of the systems are small, and no one system has lowest error rates in all focus conditions.

Three sites performed both the PE and UE tests. In [4], comparisons between the PE and UE systems for the same site indicated that the UE test condition was more difficult.

9. DISCUSSION

Not surprisingly, the same wide fluctuations in error rate observed in the 1995 tests occurred for all systems in the 1996 tests. Figure 3 shows the turn-by-turn word error rate for the CNN (commercial) news broadcast, showing three unscored regions corresponding to commercial breaks, for a typical system. Very brief "spikes" in the word error rate at, or exceeding, 100% generally correspond to the presence of background music, typically in the transitions between program segments.

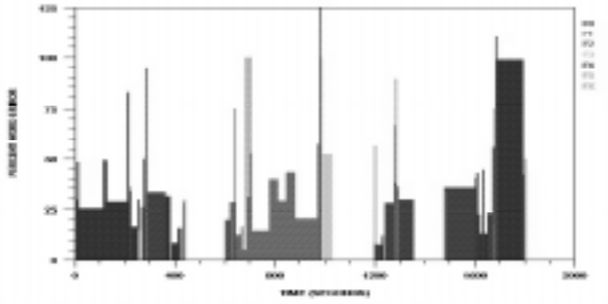

Figure 3: 1997 Error Rates vs Time

Figure 4 shows, in the form of a bar-graph with bar widths proportional to the amount of material in a given focus condition, word error rates for each focus condition for a typical system.

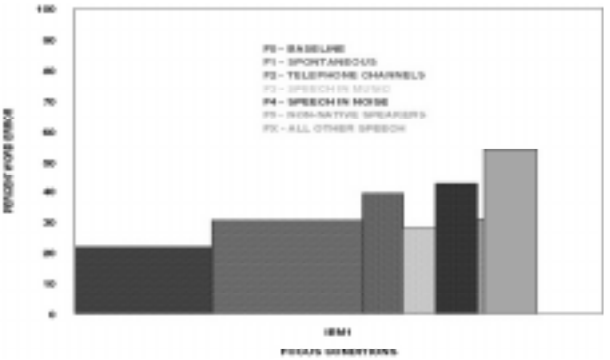

Figure 4: 1997 Focus Conditions

The F0 "baseline" material included a lengthy portion of a weather forecast from a CNN TV broadcast, although there was some discussion within our research community regarding the appropriateness of categorizing this as "planned" vs. "spontaneous" speech". Note that for this test set, the largest amount of material in any one focus condition occurred for the F1 "spontaneous speech" condition (graphically depicted in Figure 1). Because the test set yielded a small amount of F3 and F4 data, it is difficult to make a conclusion regarding the technology's effectiveness for these conditions.

Table 2 shows that for F3 the range in word error rates is from 23.3% to 50.0%, and for F4 the range is from 29.1% to 81.6%. These results serve to indicate the wide range of variability in performance that these materials produce for different systems. Highest error rates are reported for the FX focus condition consisting of combinations of conditions (e.g., spontaneous speech with background music), comprising 11.6% of the test set material. Measured word error rates ranged from approximately 50% to 70% for this focus condition.

The 1995 tests were described as "dry run" tests since no site had any experience with radio broadcast materials. The 1996 tests have been referred to as "preliminary" benchmark tests since there was no precedent for the inclusion of both radio and TV news broadcasts in NIST-administered benchmark tests within the DARPA CSR research community. Precedents now having been established, additional benchmark tests using broadcast news materials tests are being planned for November, 1997.

REFERENCES

[1] Pallett, D.S. et al., "1995 HUB-4 'Dry Run' Broadcast Materials Benchmark Tests", in *Proceedings of the Speech Recognition Workshop, February 18-21, 1996*. Distributed by Morgan Kaufmann Publishers, Inc., ISBN 1-55860-422-7.

[2] Graff, D., "The 1996 Broadcast News Speech and Language-Model Corpus", in *Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997*.
Note: "In Press" as of May, 1997. This Proceedings is to be made available in 3 formats: (1) conventional paper copy, to be distributed by Morgan Kaufmann Publishers, Inc., (2) on CD-ROM media, and (3) an on-line Web version, at a NIST web-site to be determined. For information about availability, contact david.pallett@nist.gov.

[3] Garofolo, J.S., Fiscus, J.G., and Fisher, W.M., "Design and preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora", in *Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997*.

[4] Pallett, D.S., Fiscus, J.G., and Przybocki, M.A., "1996 Preliminary Broadcast News Benchmark Tests", in *Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997*.