The TREC Spoken Document Retrieval Track: A Success Story

John S. Garofolo, Cedric G. P. Auzanne, Ellen M. Voorhees
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8940
Gaithersburg, MD 20899-8940
USA
{john.garofolo, cedric.auzanne, ellen.voorhees}@nist.gov

Abstract

This paper describes work within the NIST Text REtrieval Conference (TREC) over the last three years in designing and implementing evaluations of Spoken Document Retrieval (SDR) technology within a broadcast news domain. SDR involves the search and retrieval of excerpts from spoken audio recordings using a combination of automatic speech recognition and information retrieval technologies. The TREC SDR Track has provided an infrastructure for the development and evaluation of SDR technology and a common forum for the exchange of knowledge between the speech recognition and information retrieval research communities. The SDR Track can be declared a success in that it has provided objective, demonstrable proof that this technology can be successfully applied to realistic audio collections using a combination of existing technologies and that it can be objectively evaluated. The design and implementation of each of the SDR evaluations are presented and the results are summarized. Plans for the 2000 TREC SDR Track are presented and thoughts about how the track might evolve are discussed.

1.0 TREC

The National Institute of Standards and Technology sponsors an annual Text REtrieval Conference (TREC) that is designed to encourage research on text retrieval for realistic applications by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing results (Voorhees, et al., 2000). The conference, however, is only the tip of the iceberg. TREC is primarily an evaluation-task-driven research program. Each TREC research task culminates in a common evaluation just prior to the conference. The results of the evaluations are published by NIST in the TREC workshop notebook and conference proceedings. The sites participating in the evaluations meet at TREC to discuss their approaches and evaluation results and plan for future TREC research tasks.

In recent years the conference has contained one main task and a set of additional tasks called tracks. The main task investigates the performance of systems that search a static set of documents using new questions. This task is similar to how a researcher might use a library---the collection is known but the questions likely to be asked are not known. The tracks focus research on problems related to the main task, such as retrieving documents written in a variety of languages using questions in a single language (cross-language retrieval), retrieving documents from very large (100GB) document collections, and retrieval performance with humans in the loop (interactive retrieval). Taken together, the tracks represent the majority of the research performed in the most recent TRECs, and they keep TREC a vibrant research program by encouraging research in new areas of information retrieval. The three most recent TRECs (TREC-6 – TREC-8) have also included a Spoken Document Retrieval (SDR) track.

2.0 Spoken Document Retrieval

The motivation for developing technology that can provide access to non-textual information is fairly obvious. Large multi-media collections are already being assembled. The explosive growth of the Internet has enabled access to a wealth of textual information. However, access to audio information, and specifically spoken audio archives is pitifully limited to audio which has been manually indexed or transcribed. It is true that commerical human-generated transcripts are now available for many radio and television broadcasts, but a much greater body of spoken audio recordings (untranscribed legacy radio and television broadcasts, recordings of meetings and conferences, classes and seminars, etc.) remains virtually inaccessible. The TREC Spoken Document Retrieval (SDR) track has been created to begin to address these problems.

SDR provides content-based retrieval of excerpts from archives of recordings of speech. It was chosen as an area of interest for TREC because of its potential use in navigating large multi-media collections of the near future and because it was believed that the component speech recognition and information retrieval technologies would work well enough for usable SDR in some domains. SDR technology opens up the possibility of access to large stores of previously unsearchable audio archives and paves the way for the development of access technologies to multimedia collections containing audio, video, image, and other data formats. (Voorhees et. al., 1997a)

In practice, SDR is accomplished by using a combination of automatic speech recognition and information retrieval technologies. A speech recognizer is applied to an audio stream and generates a time-marked transcription of the speech. The transcription may be phone- or word-based in either a lattice (probability network), n-best list (multiple individual transcriptions), or more typically, a 1-best transcript (the most probable transcription as determined by the recognizer). The transcript is then indexed and searched by a retrieval system. The result returned for a query is a list of temporal pointers to the audio stream ordered by decreasing similarity between the content of the speech being pointed to and the query (Garofolo et al., 1997b). A typical SDR process is shown in Figure 1.

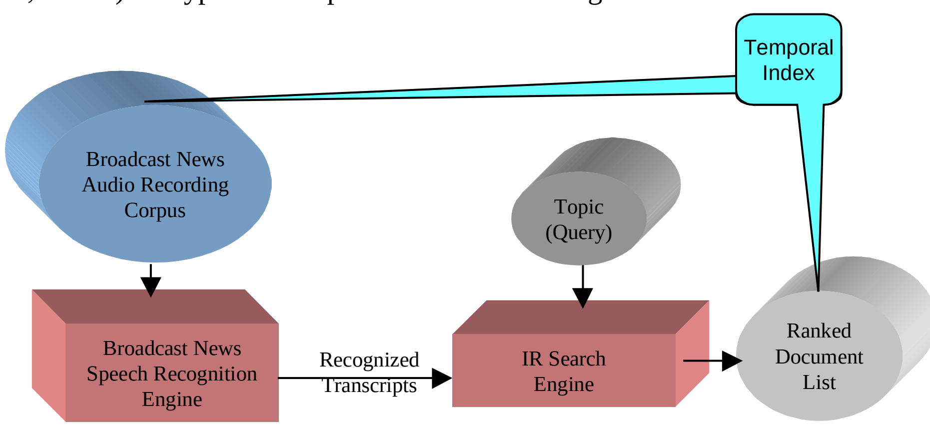

Figure 1: Typical SDR Process

3.0 TREC SDR Background

In 1996, an evaluation of retrieval using the output of an optical character recognizer (OCR) was run as a "confusion" track in TREC-5 to explore the effect of OCR errors on retrieval (Kantor, et al., 2000). This track showed that it was possible to implement and evaluate retrieval on "corrupted" text. After implementing this track, NIST and members of the TREC community thought it would be interesting to implement a similar experiment using automatic speech recognition (ASR).

During the 1996 TREC-5 workshop, researchers from NIST and the TREC community led by Karen Spärck Jones from the University of Cambridge met to discuss the possibility of applying information retrieval techniques to the output of speech recognizers. While the NIST Natural Language Processing and Information Retrieval Group had been supporting the evaluation of retrieval technologies under the auspices of TREC, the NIST Spoken Natural Language Processing Group had been working with the DARPA automatic speech recognition (ASR) community in evaluating speech recognition technology on radio and television broadcast news. The broadcast news evaluation task had accelerated progress in the recognition of real data and it seemed that the technology was producing transcripts with reasonable enough accuracy for investigation of downstream application uses such as SDR. The DARPA ASR community also had access to a 100-hour corpus of broadcast news recordings collected by the Linguistic Data Consortium (LDC) for ASR training (Graff et al., 1996) that for the first time provided a data collection which might be sufficiently large for SDR.

The NIST Spoken Natural Language Processing Group and Natural Language Processing and Information Retrieval Group joined forces to develop a plan for the creation of a research track within TREC to investigate the new hybrid technology. The primary goal of the track would be to bring the speech and information retrieval communities together to promote the development of SDR technologies and to track progress in their development. The track would also foster research on the development of large-scale, near-real-time, continuous speech recognition technology as well as on retrieval technology that is robust in the face of input errors. More importantly, the track would provide a venue for investigating hybrid systems that may be more effective than simple stove-pipe combinations. Thus, the track would also encourage cooperation and synergy between groups with complementary speech recognition and information retrieval expertise.

4.0 TREC-6 SDR: Known Item Retrieval

4.1 Evaluation Design

The first year for the SDR Track was truly one of getting the speech and IR communities together and exploring the feasibility of implementing and evaluating SDR technology. Toward that end, the TREC-6 SDR evaluation was designed for easy entry and straight-forward implementation. Since it would be the first common evaluation of SDR technology, the evaluation itself was also considered to be experimental. While the main TREC task was focussing on ad-hoc retrieval of multiple relevant documents from single topics, we decided that the first SDR Track should employ a *known-item* retrieval task which simulates a user seeking a particular, half-remembered document in a collection. The goal in a known-item retrieval task is to generate a single correct document for each topic rather than a set of relevant topics as in an ad-hoc task. This approach simplified the topic selection process and eliminated the need for expensive relevance assessments. It was also thought at the time that an SDR ad-hoc retrieval task might produce results too poor to evaluate and would discourage participation (Voorhees, et al., 1997a).

Early on we decided that the evaluation should measure not only the end-to-end effectiveness of SDR systems, but the individual ASR and IR components as well. To that end, the evaluation included several complementary runs – all using the same set of topics, but with different sets of transcriptions of the broadcast news recordings in the test collection:

Reference retrieval using "perfect"[1] human-transcribed reference transcriptions
Baseline retrieval using "given" IBM ASR-generated transcriptions
Speech retrieval using the recordings themselves, requiring both ASR and IR components

The Reference run permitted the evaluation of the overall effectiveness of the retrieval algorithms on a spoken language collection while removing ASR as a factor. Likewise, the Baseline condition permitted the comparison of the effectiveness of retrieval algorithms on the same errorful ASR-produced transcripts. Finally, the Speech run permitted the evaluation of full end-to-end SDR performance.

The Reference transcripts which were contributed by the LDC were formatted in Hub-4-style UTF format files – one for each broadcast (Garofolo, et al., 1997a). The Baseline recognizer transcripts were contributed by IBM (Dharanipragada et al., 1998). The Baseline and shared recognized transcripts were

[1] Human transcripts are not actually perfect. Hub-4 training quality transcripts are generally believed to contain 3 – 4% WER.

stored in SGML-formatted files which included story boundaries and a record for each word including start and end times. The broadcast recordings were digitally sampled (16-bit samples, linear-PCM encoded, 16-KHz. sampling rate) using a single monophonic channel and stored in NIST SPHERE-formatted files.

This componentized approach served two purposes: First, it allowed different ASR and IR sites to join together to create pipelined systems in which the components could be mixed, matched, and separately evaluated. It also permitted retrieval sites without access to ASR systems to participate in a limited way by implementing only the Reference and Baseline retrieval tasks. The participation level for sites implementing both recognition and retrieval was deemed *Full SDR* and the participation level for sites implementing retrieval only was deemed *Quasi-SDR.* Although artificial, to simplify implementation and evaluation, sites would be given human-annotated story boundaries with story ID's for all test conditions. This permitted a simplified document-based approach to implementation and evaluation.

NIST developed 47 test topics – half designed by the NIST NLPIR Group to exercise classic IR challenges. The other half were designed by the SNLP Group to exercise challenges in the speech recognition part of the problem. Half of the "speech" topics were designed to target stories with "easy-to-recognize" speech (scripted speech recorded in studio conditions with native speakers and no noise or music in the background). The other half of the speech topics were designed to target stories with "difficult-to-recognize" speech (unscripted speech, speech over telephone channels, non-native speakers, and speech with noise or music in the background). The variety of topics would permit us to examine in more detail the effect of speech recognition accuracy on retrieval performance.

We found several important differences between broadcast news stories and document-based IR collections. First, the broadcast news stories were extremely short with regard to number of words. The TREC-6 SDR collection had an average number of 276 words per story with most stories containing 100 words or less. Full-text IR collections tend to have documents with many more words – usually an order of magnitude larger. Further about 1/3 of the stories in the SDR collection were annotated as "filler" -- non-topical transitional material. We filtered the collection to remove commercials, sports summaries, weather reports, and untranscribed stories. However, we decided to leave the filler segments in the test collection to keep it as large as possible. The final filtered broadcast news collection had only 1,451 stories. Although the collection represented a sizable corpus for speech recognition (previous test corpora were less than 3 hours), it was pitifully small for retrieval testing – at least 2 orders of magnitude smaller than current IR test collections.

The test specifications and documentation for the TREC-6 SDR track are archived at http://www.nist.gov/speech/sdr97.txt.

4.2 Test Results

The test participants were given 3 months to complete the evaluation. Thirteen sites or site combinations participated in the first SDR Track. Nine of these performed Full SDR : AT&T, Carnegie Mellon University, Claritech (with CMU ASR), ETH Zurich, Glasgow University (with Sheffield University ASR), IBM, Royal Melbourne Institute of Technology, Sheffield University, and University of Massachusetts (with Dragon Systems ASR). The remaining 4 sites performed Quasi SDR: City University of London, Dublin City University, National Security Agency, and University of Maryland. (See TREC-6 SDR participant papers)

Since the goal of the track was to evaluate retrieval performance, there was no formal evaluation of recognition performance. However, Full SDR sites were encouraged to submit their 1-best transcripts so that NIST could examine the relationship between recognition performance and retrieval accuracy. The word error rate for the IBM Baseline recognizer was 50.0% (Dharanipragada et al., 1998). The mean story word error rate was a bit lower at 40%. The mean story word error rate for the other measured recognizers fell between 35% and 40%. These error rates were substantially higher than those obtained in the Hub-4 ASR tests. This difference was primarily due to three factors: The transcriptions used for scoring SDR ASR performance were created as ASR training material and had not been put through the rigorous verification that NIST employs for its Hub-4 evaluation test data. Likewise, a generic SCLITE orthographic mapping file was used. The orthographic mapping file maps alternate representations of certain words and contractions to a common format prior to scoring. A custom version of this file is created for each Hub-4 test set to minimizes the number of alternative representation confusion errors. Finally, in order to process the 50-hour collection, several sites chose to use faster, less accurate recognizers than were used in the Hub-4 tests.

Initially, we believed that the retrieval results for the SDR Track would be quite poor. Therefore, we devised scoring metrics such as *Mean Rank When Found* and *Mean Reciprocal Rank* which gave systems partial credit for finding target stories at lower ranks (Voorhees, et al, 1997a). However, we were happily surprised to find that the systems performed quite well. So well, in fact, that we chose to use *Percent Retrieved at Rank 1* as our primary metric (Garofolo, et al, 1997b). Retrieval rates were very high for the Reference transcript condition and most sites showed only a small degradation for retrieval using their own recognizers. There was generally higher degradation in retrieval using the Baseline recognizer transcripts due to its high error rate and high number of *out-of-vocabulary* (OOV) words. The results of the evaluation for all three retrieval conditions are shown in Figure 2.

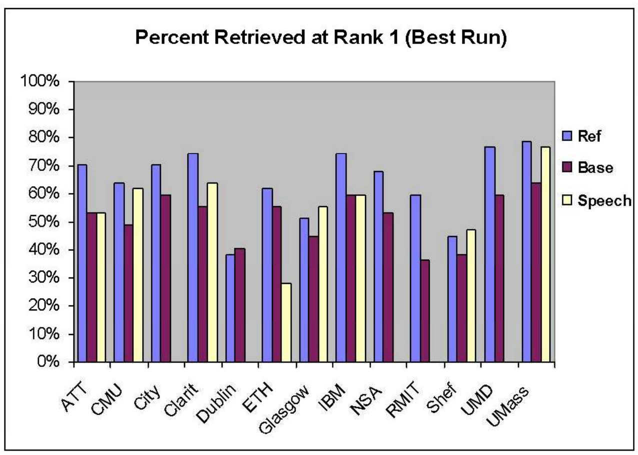

Figure 2: TREC-6 SDR Retrieval rate at rank 1 for all systems and modes (best run)

For Percent Retrieved at Rank 1, the best performance for all three test conditions was achieved by the University of Massachusetts System (with Dragon Systems recognition for Full SDR) which obtained a retrieval rate of 78.7% for the Reference condition, 63.8% for the Baseline recognizer condition, and 76.6% for the Speech condition (Allan et al, 1997). In fact, the UMass system missed only one more topic on the Speech condition than it did on the Reference condition.

An analysis of errors across systems for particular topics (Figure 3) showed that, in general, the "Easy to Recognize" topic set yielded the best performance for all 3 evaluation conditions while the "Difficult to Recognize" topic set yielded substantially degraded performance. However, the "Difficult Query" topic subset yielded even greater performance degradation. It is interesting to note that systems also had difficulty in retrieving stories for the "Difficult to Recognize" topic subset from the Reference transcriptions – an indication that factors in transcribed speech other than recognition errors might influence retrieval performance. However, there was far too much variance from the topic effect to make any sweeping conclusions.

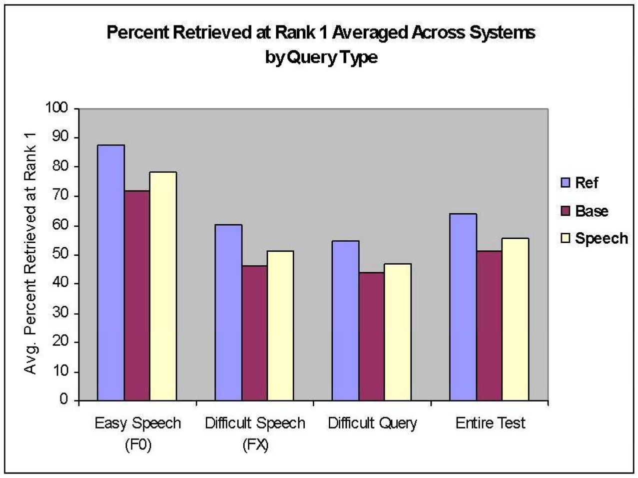

Figure 3 : TREC-6 SDR Percent Retrieval at Rank 1 averaged across systems by topic subset

To further examine the effect of recognition error rate on retrieval, we examined performance using the Baseline recognizer results. For each topic, we sorted the mean rank at which the retrieval systems found the target story against the word error rate for that story (Figure 4). The sorting appears to show an increasing trend toward poorer retrieval performance as recognition errors increase.

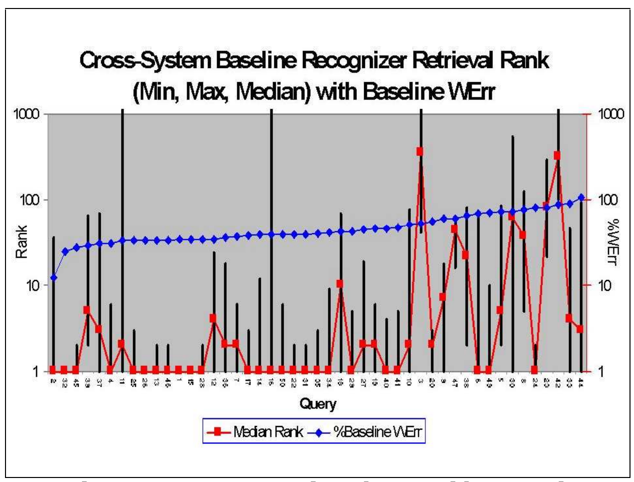

Figure 4 : TREC-6 Baseline condition mean retrieval rank sorted by Baseline Recognizer story word error rate

Interestingly, the same plot for retrieval for the Reference transcripts shows a similar trend (Figure 5) indicating that stories that are difficult to recognize may also be innately difficult to retrieve – even when recognized perfectly. One hypothesis is that the complexity of the language within the more difficult-to-recognize stories is greater than that of the more easy-to-recognize stories.

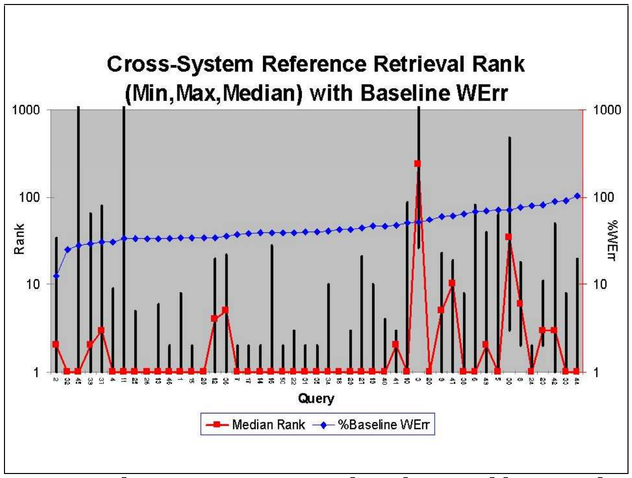

Figure 5 : TREC-6 Reference condition mean retrieval rank sorted by Baseline Recognizer story word error rate

A statistical analysis of variance showed that we had too little data to eliminate a large proportion of confounding unexplained factors (Garofolo, et al., 1997b). A future evaluation which would provide multiple recognizer transcript sets which all retrieval sites would run against would help to clarify the relationship between recognition and retrieval performance.

4.3 Conclusions

The first SDR evaluation showed us that we could successfully implement an evaluation of SDR technology and that existing component technologies worked well on a known-item task with a small audio collection. However, the test participants all agreed that the test collection would have to be enlarged by at least an order of magnitude before any “real” performance issues would surface. It was also agreed that the known-item task provided insufficient evaluation granularity. For this evaluation, it seemed that retrieval performance played a much more significant role in overall SDR performance than recognition performance. However, it was difficult to make any conclusions given the limited evaluation paradigm and collection.

5.0 TREC-7 SDR : Ad Hoc Retrieval

5.1 Evaluation Design

In 1998, for TREC-7, we set out to address some of the inadequacies in the TREC-6 SDR Track. We still did not have access to a large enough audio collection for true retrieval evaluation, but we were able to double the size of the SDR collection using an additional broadcast news corpus collected by the LDC for Hub-4 ASR training. More importantly, though, we decided to give up the known item retrieval paradigm and implement a classic TREC ad-hoc retrieval task.

In an ad hoc retrieval test, systems are posed with topics and attempt to return a list of documents ranked by decreasing similarity to the topic. The documents are then evaluated for relevance by a team of human assessors. In TREC, to keep the evaluation tractable, NIST pools the top *N* documents output by all of the evaluated systems and judges only those documents. Therefore, systems get evaluated over all documents, but only some documents are judged. Although not exhaustive, this approach assumes that with enough different systems, all of the relevant documents will be included in the pool. The traditional TREC ad-hoc track provided several forms of information for each topic: A title, a short query form -- usually a single sentence or phrase, and a descriptive narrative giving rules for judging relevance. Given the limited size of the SDR collection, we decided to simplify the SDR topics to a single short form. We also required that all runs had to be fully automatic.

The TREC-7 SDR test collection contained 87 hours of audio with 2,866 usable stories after filtering and a similar mean and median story length as compared to the TREC-6 collection. As in TREC-6, participants were given human-annotated story boundaries and story IDs. This removed story-boundary detection from the technical challenge, but permitted NIST to use the standard TREC document-based TREC_EVAL scoring software to evaluate the results of the test. A team of 3 NIST TREC assessors created 23 test topics (averaging 14.7 words in length) for the collection. The following are two of the test topics they created:

> *Find reports of fatal air crashes.* (Topic 62)
>
> *What economic developments have occurred in Hong Kong since its incorporation in the Chinese People's Republic?* (Topic 63)

To more accurately examine the effect of recognition performance on retrieval, we decided to add a new optional evaluation condition, *Cross Recognizer Retrieval*, in which retrieval systems would run on other sites' recognized transcripts. This would permit us to more tightly control for the recognizer effect in our analyses as well as provide us with more information regarding the relationship between recognizer performance and retrieval performance. We therefore encouraged all sites running 1-best recognition to submit their recognizer transcripts to NIST for sharing with other participants. To permit sites to explore the effect of using different recognizers, we permitted each Full SDR site to run retrieval on both a primary (S1) and secondary (S2) recognizer.

For the Baseline recognizer, NIST created a local instantiation of the Carnegie Mellon University SPHINX-III recognizer. Since SPHINX-III ran in nearly 200 times real time on NIST's UNIX-based workstations, NIST realized that it would take nearly two years of computation to complete a single recognition pass over the 87-hour collection. NIST learned of inexpensive clusters of PC-LINUX-based systems being used by NASA in its BEOWULF project (BEOWULF, 1997) and set out to create a cluster-based recognition system. The final system incorporated a scheduling server and 40 computational nodes. Given the cluster's enormous computational power, to further enrich the spectrum of recognizers in the evaluation, NIST chose to create two Baseline recognizer transcript sets. One set (B1) was created using an "optimal" version of the SPHINX recognizer and benchmarked at 27.1% word error rate on the Hub-4 '97 test set (Pallett, et al., 1998) and at 33.8% on the SDR test collection. This enabled us to for the first time benchmark the difference in performance for the same recognizer running both Hub-4 and SDR ASR tests. A second set (B2) was created using lowered pruning thresholds and benchmarked at 46.6% word error rate for the SDR collection.

As in TREC-6, Full SDR sites were required to implement the Reference, Baseline, and Speech input retrieval conditions and the Quasi SDR sites were required to implement only the Reference and Baseline retrieval conditions.

The test specifications and documentation for the TREC-7 SDR track are archived at http://www.nist.gov/speech/sdr98/sdr98.htm.

5.2 Test Results

The TREC-7 SDR participants were given 4 months to implement the recognition portion of the task. They were then given one month to implement the required retrieval tasks and an additional month to implement the optional Cross Recognizer retrieval task. The sites were not restricted in the hardware or number of processors they could apply in implementing the evaluation.

Eleven sites or site combinations participated in the second SDR Track. Eight of these performed Full SDR: AT&T [ATT], Carnegie Mellon University Group 1 [CMU1], University of Cambridge [CUHTK], DERA [DERA], Royal Melbourne Institute of Technology [MDS], Sheffield University [SHEF], The Netherlands Organization - TPD TU-Delft [TNO], and University of Massachusetts (with Dragon Systems ASR) [UMass]. The remaining 3 sites performed Quasi SDR : Carnegie Mellon University Group 2 [CMU2], National Security Agency [NSA], and the University of Maryland [UMD]. (See TREC-7 SDR participant papers)

In addition to the two NIST Baseline recognizers, 1-best transcripts for 6 additional recognizers were submitted to NIST for scoring and sharing in the Cross Recognizer retrieval condition. The recognizers covered a wide range of error rates and provided a spectrum of material for the Cross Recognizer retrieval condition. Figure 6 shows the word error rate and mean story word error rate for each of the submitted recognizer transcripts.

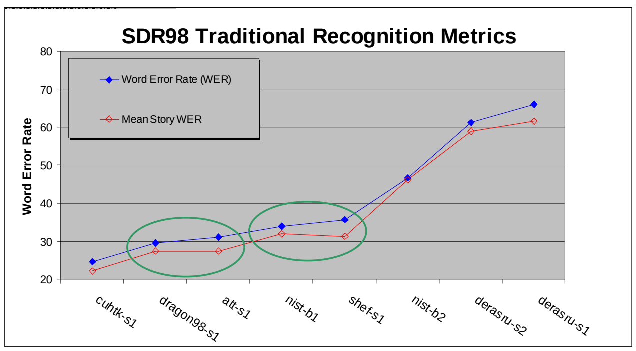

Figure 6: TREC-7 SDR Test set word error rate (WER) and mean story word error rate (SWER) for submitted recognized transcripts with cross-system significance at 95% for SWER

The best recognition results were obtained by the University of Cambridge HTK recognition system with a 24.6% test set word error rate and a 22.2% mean story word error rate (Johnson, et al., 1998). The circled mean story word error rate points were not considered to have statistically different performance. While the SDR ASR error rates were still significantly higher than Hub-4, in general, error rates were significantly improved from the previous year – even at the faster speeds required to recognize the larger test collection.

Each retrieval run was required to produce a rank-ordered list of the ID's for the top 1000 stories for each topic. The top 100 IDs from each of these lists were then merged to create the pools for human assessment. The 3 TREC assessors read the reference transcriptions for each of the topic pool stories to evaluate the stories for relevance. All of the retrieval runs were then scored using the standard TREC_EVAL text retrieval scoring software. As in other TREC ad hoc tasks, the primary retrieval metric for the SDR evaluation was mean average precision (MAP) which is the mean of the average precision scores for each of the topics in the run. The average precision is equivalent to the area underneath the uninterpolated recall-precision graph (Voorhees, et al., 1998).

In all, the TREC-7 SDR Track contained 6 retrieval conditions :

Reference (R1): retrieval using Human (closed-caption-quality) reference transcripts
Baseline-1 (B1): retrieval using NIST (CMU SPHINX) ASR transcripts
Baseline-2 (B2): retrieval using NIST (CMU SPHINX) "sub-optimal" ASR transcripts

Speech-1 (S1): retrieval using participant's own recognizer
Speech-2 (S2): retrieval using participant's own secondary recognizer
Cross Recognizer (CR) : retrieval using other participants' recognizer transcripts

The results for each of the required test conditions: Reference (R1), Baseline-1 (B1), Baseline-2 (B2), Speech-1 (S1) and Speech-2 (S2) are shown in Figure 7. Full SDR participants were required to implement the R1, B1, B2, and S1 retrieval conditions. Quasi SDR participants were required to implement the R1, B1, and B2 retrieval conditions.

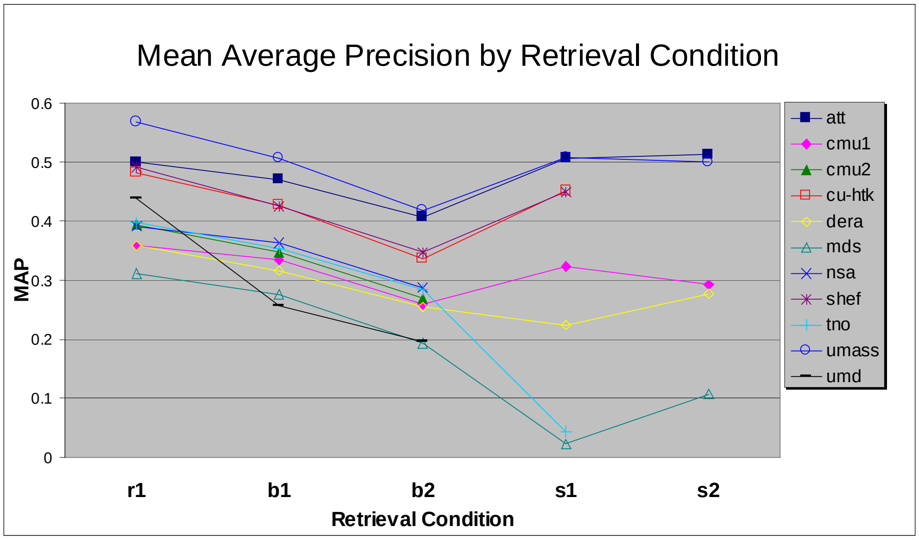

Figure-7: TREC-7 SDR Mean Average Precision (MAP) for required retrieval conditions

For all retrieval conditions except S2, the University of Massachusetts system (Allan, et al., 1998) achieved the best mean average precision. Most systems performed surprisingly well for the recognizer-based conditions. Even more surprising, AT&T's S2 run (the best recognizer-based run in the evaluation) outperformed its R1 run. AT&T attributed this excellent performance to a new approach they implemented for document expansion using contemporaneous newswire texts which they employed for their S1/S2 runs but not for their R1 run (Singhal, et al., 1998).

The most interesting condition for TREC-7 SDR was the cross recognizer retrieval (CR) condition in which participating systems ran retrieval on the 6 submitted recognizer-produced transcript sets in addition to the human Reference and B1/B2 recognizer transcript sets. This experiment gave us 9 recognition/retrieval data points to examine the effect of recognition performance on retrieval performance. Four sites (University of Cambridge, DERA, Royal Melbourne Institute of Technology [MDS], and Sheffield University) participated in the CR experiment. Using the mean story word error rate (SWER) ASR metric and the mean average precision (MAP) retrieval metric, we plotted the recognition/retrieval performance curve for each of the four systems (Figure 8).

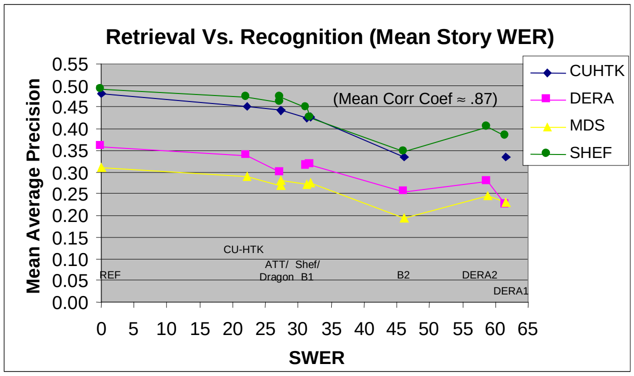

Figure 8: TREC-7 SDR Cross Recognizer results: mean average precision vs. mean story word error rate

The figure shows a gentle, but fairly linear drop-off in MAP for recognition transcripts with increasing SWER. We calculated the correlation coefficient for the metrics to determine how well SWER correlated with retrieval performance. The average correlation coefficient for the 4 systems was .87 – a significant correlation.

We explored several other word-error-rate-based metrics to see if we could find an even better predictor for retrieval performance. Our hypothesis was that such a metric would be useful in developing ASR systems for retrieval purposes. We explored metrics which used IR methods to filter out unimportant words for retrieval: *stop-word-filtered word error rate* and *stemmed stop-word-filtered word error rate* (Garofolo, et al., 1998). Surprisingly, however, these metrics turned out to be only slightly more correlated with mean average precision than word error rate. Other effective approaches to IR-customized ASR scoring using the TREC SDR data have been explored and reported by Johnson (1999) and Singhal (1999).

While we were implementing the TREC-7 SDR track, we were also administering a first evaluation in Named Entity (NE) tagging using broadcast news. The NE evaluation involved identification of people, locations, and organizations in broadcast news ASR transcripts (Przybocki, et al., 1999). To our fortune, GTE/BBN had hand-annotated the same data we used in the SDR evaluation with Named Entity tags (Miller, et al., 1999). Our hypothesis was that these named entities would identify most of the key content-carrying words in our spoken documents and that if we focussed our ASR metric on these words, we would obtain a better predictor of retrieval performance than by measuring the error rate of all words. We re-scored the ASR systems using the named entity word error rate and plotted the ASR metric against the mean average precision as we had done with mean story word error rate (Figure 9).

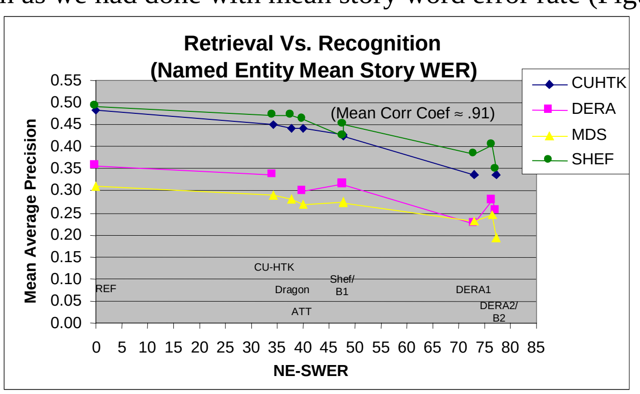

Figure 9: TREC-7 SDR Cross Recognizer results: mean average precision vs. named entity mean story word error rate

The plot showed a nearly linear relationship between named entity ASR performance and retrieval performance with a mean correlation coefficient of .91 across the systems. Most significantly, the plot more accurately positioned the problematic NIST B2 recognizer which had systematically-increased errors in longer (probably more-content-carrying) words. For all the systems, the named-entity-based metric showed a higher correlation with mean average precision than word error rate alone (Garofolo, et al., 1998). Other things being equal, this finding tells us that an ASR system which recognizes named entities most accurately will provide the best input for retrieval.

5.3 Conclusions

For TREC-7, we learned that we could successfully implement and evaluate an ad hoc SDR task. With the new Cross Recognizer condition, we were able to begin to investigate the relationship between recognition performance and retrieval performance. We found a near-linear relationship between word error rate and mean average precision and we found that recognition content-word-based word error metrics such as named entity word error rate provided even better predictors of retrieval performance than word error rate alone. Although twice the size of it's predecessor in number of stories, our 87-hour collection was still too far too small to make conclusions about the usefulness of the technology. Further, we were still evaluating systems using artificial human-annotated story boundaries.

6.0 TREC-8 SDR : Large Audio Collection

6.1 Evaluation Design

In 1998, the Linguistic Data Consortium began collecting a large radio and television corpus for the DARPA Topic Detection and Tracking (TDT) program. In contrast to most TREC tracks[2], the TDT program, is concerned with detecting and processing information from a continuous stream as it occurs in an *online* manner (Fiscus, et al., 1999). The TDT-2 corpus, collected to support the TDT program in 1998-99, contains news recordings from ABC, CNN, Public Radio International, and the Voice of America. With the exception of the VOA broadcasts, which began in early March, these sources were sampled evenly over a 6-month period between January and June 1998. The corpus also contains a contemporaneous newswire corpus containing articles from the New York Times and Associated Press (Cieri, et al., 1999).

With it's time-sampled broadcast news sources and parallel text corpus, the 600-hour TDT-2 corpus was also almost perfectly suited for use in the SDR Track. Unfortunately, it had no high-quality human reference transcriptions – only "closed-caption" quality transcriptions. Since the transcription quality prevented us from reasonably evaluating recognition performance over the entire collection, we selected a 10-hour randomly-selected story subset of the collection for detailed transcription by the LDC. These high-quality transcripts would permit us to perform a sampled evaluation of the ASR performance. They also permitted us to evaluate the error rate in the closed-caption-quality transcriptions themselves which we found to have roughly 14.5% WER for television closed-caption sources and 7.5% WER for radio sources which had been quickly transcribed by commercial transcription services (Fisher, 1999). These error rates are significant and the television closed caption error rates approach the error rates for state-of-the-art broadcast news recognizers.

Several SDR participants were also Hub-4 participants and intended to use their Hub-4 ASR systems which contained training data from January 1998 (which overlapped with the first month of the TDT-2 corpus.) To eliminate the possibility of training/test cross-contamination, we eliminated the January data from the SDR collection. The final collection contained 557 hours of audio collected between February

[2] The TREC Filtering track works on an online retrieval task similar to TDT.

1, 1998 and June 30, 1998. The collection contained 21,754 stories – an order of magnitude larger than the 87-hour TREC-7 SDR collection.[3]

We believe that deployed SDR systems will operate in an archive search modality. The most efficient means to implement such a system is to employ *online recognition* (in which recognition is performed on a continuous basis as audio is recorded) and *retrospective retrieval* in which the entire collection is queried after it is formed. This is in contrast to a TDT-type system which performs *online retrieval* as the audio is recognized. In both modalities, recognition should use adaptation techniques to adjust to changes in the collection language over time. Traditional Hub-4-style broadcast news recognizers employed only static pre-trained language models. If such a recognizer was used in a real time-longitudinal application, the language in the news and the fixed language model used in the recognizer would diverge, resulting in increasing error rates over time. Such recognizers are incapable of recognizing new words – words likely to be important for retrieval. Conversely, given the computational expense of performing recognition, *retrospective recognition* at the time of retrieval is impossible for realistically large collections. So, in a real SDR application where audio would be recorded over many months or years, the recognizer would have to be re-trained periodically to accommodate changes in the language and new words. To support this modality, we defined an online recognition mode which supported the use of evolving "rolling" language models in which the recognition systems could be periodically retrained over the test epoch. Full SDR sites were permitted to use either a traditional pre-trained recognition system or a continuously adaptive recognition system which used the contemporaneous newswire text from days prior to the day being recognized for adaptation. Sites were free to choose whatever retraining period or strategy they liked as long as they didn't "look ahead" in time as they performed recognition (Garofolo, et al., 1999).

Realizing that the CMU SPHINX recognizer was far too slow to recognize the TREC-8 collection, NIST set out to find a faster baseline recognizer. During 1998, NIST added a spoke to its Hub-4 broadcast news ASR evaluation in which systems had to run in 10 times real time or fast on a single processor. This spoke, dubbed *10Xrt*, encouraged the development of fast broadcast news recognizers which suffered little degradation in recognition accuracy over their 150Xrt+ cousins (Pallett, et al., 1999). GTE/BBN offered NIST a LINUX instantiation of their fast BYBLOS Rough 'N Ready recognizer (which now operated at 4Xrt) to use as a baseline in the SDR and TDT tests (Kubala, et al., 2000). BBN also gave NIST a basic language modeling toolkit to work with. Given the computational power of NIST's recognition cluster and the speed of the BBN recognizer, NIST set out to create 2 complementary baseline recognizer transcript sets. The first set (B1) used a traditional Hub-4 fixed language model. The B1 recognizer benchmarked at 24.7% WER on the Hub-4 '97 test set, 23.4% WER on the Hub-4 '98 test set, and 27.5% WER on the SDR-99 10-hour subset. NIST then created an adaptive "rolling" language model version (B2) that used the SDR contemporaneous newswire texts for periodic look-back language model training. Details regarding the B2 recognizer are provided in Auzanne, et al. (2000). The B2 system benchmarked at 26.7% WER on 10-hour SDR-99 subset. This difference in performance might seem insignificant. However, NIST statistical tests showed that it is significantly different than the B1 recognizer. Further, the small decrease in word error belies a more significant decrease in the out-of-vocabulary (OOV) rate of the recognizer. The OOV rate is the percentage of test set words which are not included in the recognizer's vocabulary and which, therefore, can never be correctly recognized. The OOV rate for the fixed B1 recognizer was 2.54%. The OOV rate for the adaptive B2 recognizer was 1.97% -- a 22.4% relative improvement.

In addition to the Reference, Baseline, Speech, and Cross Recognizer retrieval conditions used in TREC-7, an optional story boundaries unknown (SU) condition was added for TREC-8. This condition

[3] The difference in story density is explained by the large proportion of short CNN stories in the TREC-8 collection. The average story length in the TREC-8 collection is only 169 words.

permitted sites to explore SDR where they had to operate on whole broadcasts with no knowledge of human-annotated topical boundaries. This condition more accurately represented the real SDR application challenge. A new ad-hoc paradigm had to be created to support the SU condition since it was not document based as in previous evaluations. The natural unit for audio recordings is time rather than documents or words. Therefore, it was decided that SU systems would output a ranked list of time pointers. Given that the TDT program was already investigating technology for story segmentation, we did not want to require SDR systems to find the topical boundaries in the audio recordings. Rather, we decided to require them to emit only a single time pointing to a "hot spot" or mid-point of a topical section. This approach allowed us to map the emitted times to known stories and make use of our traditional document retrieval evaluation software. Thus, this approach focussed on a new and interesting problem while making use of the existing evaluation infrastructure and permitting some comparison between runs where story boundaries were known and runs where they weren't known. To keep the task clean, we required that Full SDR sites implementing the SU option would also be required to run their recognizers without knowledge of story boundaries. However, to make maximal use of the recognizers for the CR task, NIST devised a script to backfill the story boundaries into the SU ASR transcripts.

The new SU condition did pose some challenges for scoring. The biggest issue was how time pointers which mapped to the commercials, fillers, or the same stories should be treated. NIST decided to implement a mapping algorithm that would severely penalize the over-generation of time pointers. The pointers were first mapped to known story ID's. Duplicate story ID's, commercials, and fillers were then mapped to "dummy" ID's which would be automatically scored as non-relevant. The results were then scored as usual with TREC_EVAL. Since the story boundary known (SK) collection excluded commercials and other untranscribed segments that were included in the SU collection, direct comparisons between the two conditions would not be possible. However, this first SU evaluation would give us an idea of how difficult a technical challenge the SU condition would pose.

A team of 6 NIST assessors created the ad hoc topics for the evaluation. The goal in creating TREC topics is to devise topics with a few (but not too many) relevant documents in the collection to appropriately challenge retrieval systems. Prior to coming together at NIST, the assessors were told to review the news for the first half of 1998 and to come up with 10 possible topics each. The assessors then tested their putative topics against the Reference transcripts in the TREC-8 SDR collection using the NIST PRISE search engine. If a topic was found to retrieve 1 to 20 documents in the top 25, it was considered for inclusion in the test. Otherwise, the assessors were required to refine (broaden or narrow) or replace the topic to retrieve an appropriate number of relevant documents using PRISE. The assessors created approximately 60 topics. Topics with similar subjects or which were considered malformed were then excluded to yield the final test set containing 49 topics.

The test specifications and documentation for the TREC-8 SDR track are archived at http://www.nist.gov/speech/sdr99/sdr99.htm.

6.2 Test Results

The TREC-8 SDR participants were given approximately three and a half months to implement the recognition portion of the task and a month and a half to implement the required retrieval tasks. In order to give the participants the maximum possible amount of time to run recognition, the retrieval period overlapped the recognition period by one month. After the site's recognized transcripts were submitted to NIST, they were checked, filtered, formatted and distributed for the Cross Recognizer retrieval condition. The retrieval sites were then given 3 weeks to perform the CR task. Since NIST had limited time for assessment, only the pre-CR retrieval results were used to construct the pools for assessment, which took place in parallel with the CR test. As in TREC-7, the sites were not restricted in the hardware or number of processors they could apply in implementing the evaluation.

Ten sites or site combinations participated in the third SDR Track. Six of these performed Full SDR: AT&T [ATT], Carnegie Mellon University [CMU], University of Cambridge [CU-HTK], LIMSI [LIMSI], Sheffield University [SHEFFIELD], and Twenty One Consortium [TNO]. The remaining 4 sites performed Quasi SDR: The State University of NY at Buffalo [CEDAR], IBM [IBM], The Royal Melbourne Institute of Technology [MDS], and the University of Massachusetts [UMASS]. (See the TREC-8 participant publications)

In all, the TREC-8 SDR Track contained 11 retrieval conditions:

Reference (R1): retrieval using Human (closed-caption-quality) reference transcripts
Baseline-1 (B1): retrieval using NIST (BBN Byblos) fixed language model ASR transcripts
Baseline-2 (B2): retrieval using NIST (BBN Byblos) adaptive language model ASR transcripts
Speech-1 (S1): retrieval using site's own recognizer
Speech-2 (S2): retrieval using site's own secondary recognizer
Cross Recognizer (CR): retrieval using other site's recognizer transcripts
Baseline-1 boundaries unknown (B1U)
Baseline-2 boundaries unknown (B2U)
Speech-1 boundaries unknown (S1U)
Speech-2 boundaries unknown (S2U)
Cross Recognizer boundaries unknown (CRU)

Full SDR sites were required to run the R1, B1, and S1 retrieval conditions. Quasi-SDR sites were required to run only the R1 and B1 retrieval conditions. The B2, CR and all story boundaries unknown conditions (*U) were optional.

We benchmarked the performance of the speech recognizer transcripts contributed by Full SDR sites for sharing in the Cross Recognizer condition using the 10-hour Hub-4-style transcribed subset of the SDR collection. The summary results are shown in Figure 10.

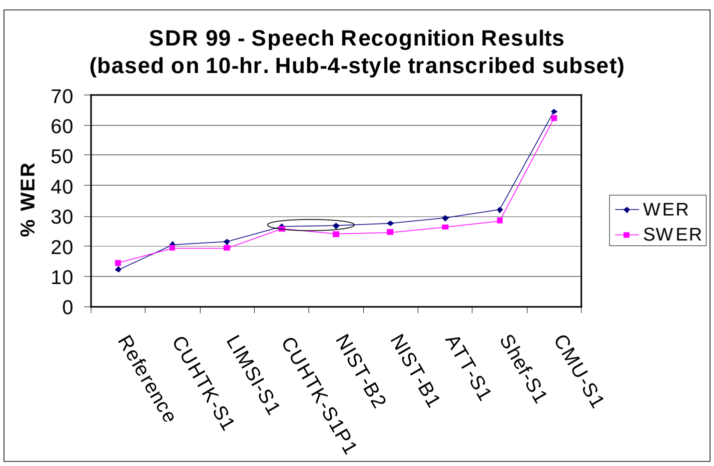

Figure 10 : TREC-8 SDR Speech Recognition Performance Results (Test Set Word Error Rate and Mean Story Word Error Rate) with cross-system significance for Word Error Rate

The word error rates were surprisingly low considering the enormous size of the test collection which was over 2 orders of magnitude larger than test sets used in Hub-4 ASR tests. The graph shows the results for both test-set word error rate and mean story word error rate. Most of the systems produced transcripts with word error rates of less than 30%. This is fairly impressive considering the speed at which the systems had to be run to process the large collection. It is also interesting to note that these scores are generally lower than the comparable scores from TREC-7 in which ASR systems were not run at such fast speeds. The best ASR results were obtained by the University of Cambridge HTK recognizer with a 20.5% WER (Johnson, et al., TREC-8 1999). With the exception of the alternative first-pass-only Cambridge System and the NIST B2 system, none of the recognizer transcripts were found to be significantly similar in performance with respect to WER by the NIST statistical significance software. The figure also shows the results of scoring the original closed-caption-style Reference transcripts against the more scrupulously transcribed Hub-4-style transcripts.

As with the speech recognition performance, overall retrieval performance was quite good. As with all TREC ad hoc tests, there was quite a bit of variation in performance for particular topics. The following sample TREC-8 SDR test topics illustrate the variation:

Topic 105: *How and where is nuclear waste stored in New Mexico?*
(.85 average MAP across all systems/runs, 7 relevant stories).

Topic 117: *If we get more income, will we save more or spend more?*
(.34 average MAP across all systems/runs, 28 relevant stories)

Topic 94: *What percentage of the population is in prison in the U. S. A. and in the E. C. countries?*
(.01 average MAP across all systems/runs, 7 relevant stories)

Figure 11 shows the results for each of the non-Cross-Recognizer retrieval conditions. The best results for the Reference and Baseline-1 recognizer retrieval conditions were obtained by the AT&T system, with a MAP of .5598 and .5539 respectively (Singhal, et al., TREC-8 1999). The best result for the Speech input retrieval condition was obtained by the University of Cambridge system with a MAP of .5529 (Johnson, et al., TREC-8 1999). Sheffield University achieved the best performance for the Baseline and Speech input story boundary unknown conditions with a MAP of .4301 and .4250 respectively (Abberley et al.,1999).

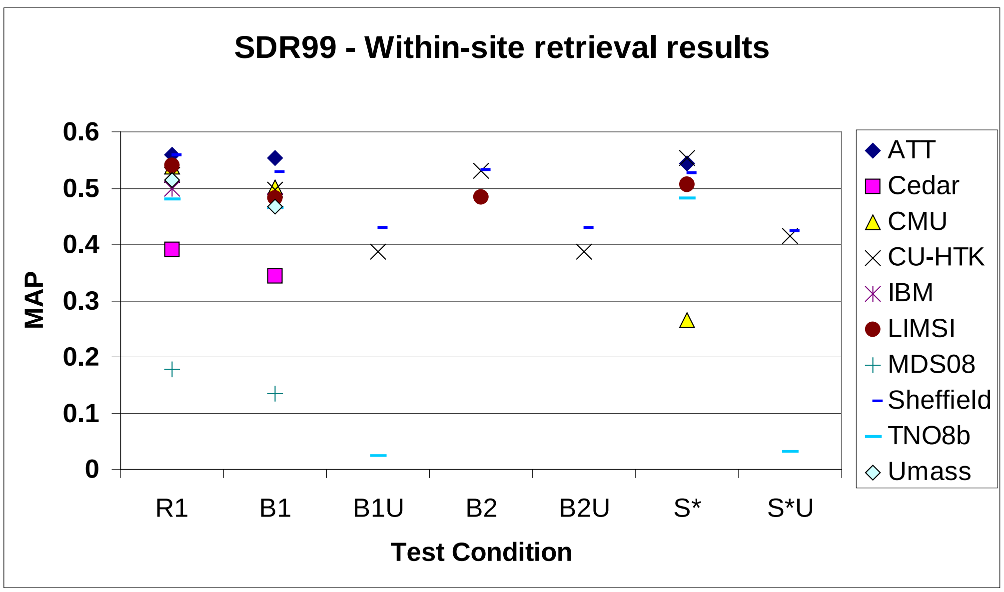

Figure 11: TREC-8 SDR Mean Average Precision (MAP) for required and non-cross-recognizer retrieval conditions

The individual test conditions were useful in contrasting the effect of binary variables such as human transcripts vs. ASR transcripts and story boundaries known vs. story boundaries unknown. However, even more interesting results are found in the Cross-Recognizer retrieval conditions which contain multiple recognition performance/retrieval performance data points with which we can examine the effect of recognition performance on retrieval performance.

Four sites participated in the story boundaries known Cross-Recognizer (CR) retrieval condition: AT&T, University of Cambridge, LIMSI, and Sheffield University. Each of these sites ran retrieval on the 8 sets of submitted recognizer transcripts. Adding the retrieval results for the closed-caption-quality Reference transcripts, this gives us 9 recognition/retrieval data points for each system. Figure 12 shows a graph of retrieval performance vs. recognition performance for the story boundaries known Cross-Recognizer retrieval condition. The CMU recognizer data point was removed since it was an extreme outlier. The graph shows that retrieval performance degrades very little for transcripts with increasing word error rates and that retrieval is fairly robust to recognition errors. Our hypothesis is that the redundancy of key words in the spoken documents permits the relevant documents to be retrieved – even when a substantial number of words are mis-recognized. For TREC-7, we assumed that this robustness was due to the small collection size and expected the recognition/retrieval performance drop-off to be much steeper for the larger TREC-8 collection. However, this does not appear to be the case. When we compare the average cross-system slope for the recognition/retrieval performance curve for TREC-7 and TREC-8, we find that they are almost identical (.0016 for TREC-8 vs. .0014 for TREC-7). Although the individual systems had different relative retrieval performance, all of the systems' slopes appears to be relatively flat. The AT&T system achieved the best CR performance and also had the most shallow recognition/retrieval performance slope (Singhal, et al., TREC-8 1999).

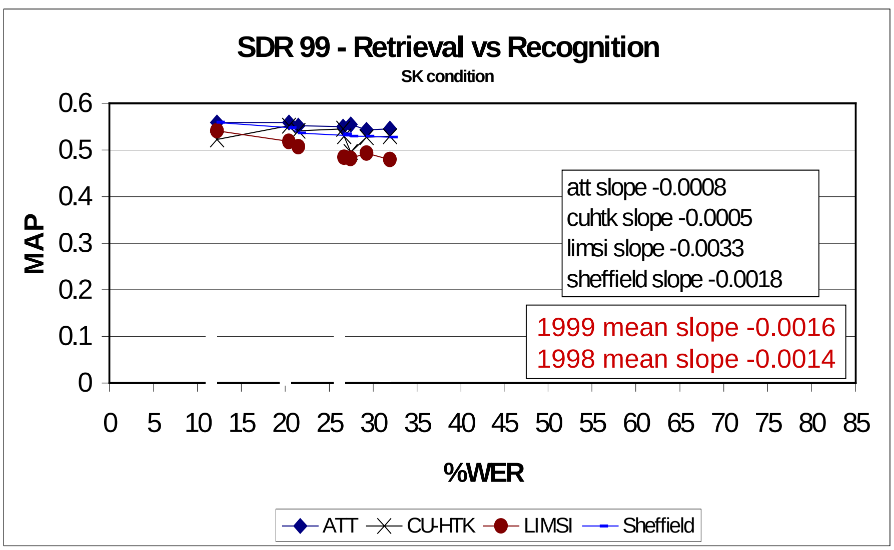

Figure 12 : TREC-8 SDR Story Boundaries Known Cross Recognizer Retrieval condition results showing Mean Average Precision vs. Word Error Rate

Three sites participated in the story boundaries unknown Cross Recognizer (CRU) retrieval condition: University of Cambridge, Sheffield University, and The Twenty One Consortium. The results of the CRU condition are shown in Figure 13.

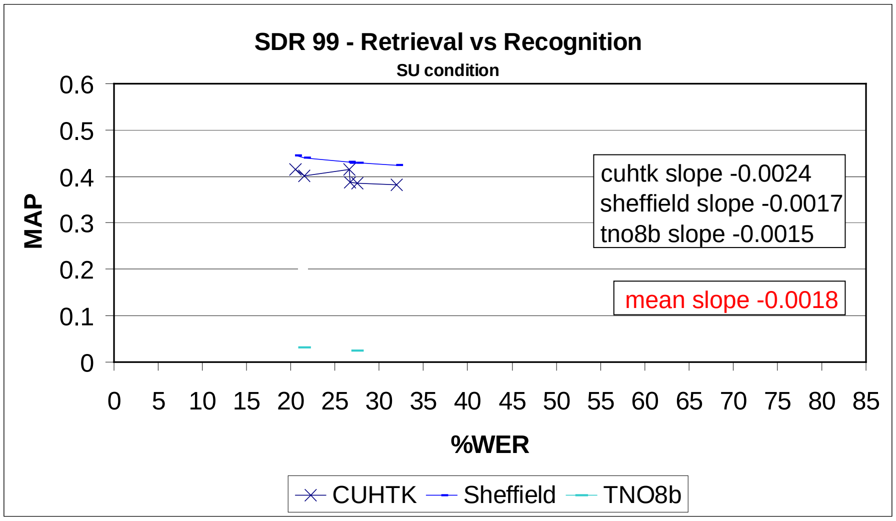

Figure 13 : TREC-8 SDR Story Boundaries Unknown Cross Recognizer Retrieval condition results showing Mean Average Precision vs. Word Error Rate

As in the story boundaries known CR condition, although the relative performance of the retrieval systems differed, their recognition/retrieval performance slopes were relatively flat with an average cross-system slope of .0018. The University of Cambridge system achieved the best CRU performance (Johnson, et al., TREC-8 1999). The CRU retrieval scores are significantly lower than the comparable CR scores, which indicates that the unknown story boundaries pose greater difficulties for the retrieval systems. Part of this difficulty is explained by the difference in test data. The story boundaries known systems used transcripts in which commercials, filler, and untranscribed segments were removed, whereas the story boundaries unknown systems had to process the entire broadcasts. It is even more difficult to compare the results given the penalization for duplicates in scoring.

6.3 Conclusions

The recognition results for the TREC-8 were extremely encouraging. We saw recognition error rates fall even as recognition systems were made faster to tackle the large TREC-8 collection. The results for the retrieval systems were also quite good. Given these factors, we can conclude that not only is the technology robust to larger spoken document collections, but that it has also improved significantly since TREC-7. We found that adaptive recognition systems can be used to more effectively recognize speech data collected over time than comparable static systems. The Cross Recognizer retrieval conditions with its multiple recognition/retrieval data points showed us that there is a near-linear relationship between recognition errors and retrieval accuracy and that the retrieval performance degradation slope for increasing recognition errors is relatively gentle. Finally, we found that SDR technology can be applied to, and evaluated for conditions in which story boundaries are unknown.

7.0 TREC-9 SDR Plans

After much discussion, the TREC SDR community has decided to stabilize the SDR track for the upcoming year with only a few minor changes. The most significant of these is that the story boundaries unknown condition will be mandatory for all participants. The same test collection will be used as in 1999, but a new set of 50 test topics will be developed. Since the story boundaries unknown condition can make effective use of audio-signal information not found in the transcriptions such as speaker changes, noise changes, volume changes, music, prosody, etc., we will encourage the development of a common non-lexical information exchange format which can be used to store and share such information. We will also encourage SDR participants this year to share this data in addition to their ASR transcripts for the cross recognizer retrieval condition.

The test specifications and documentation for the TREC-9 SDR track will be made available at http://www.nist.gov/speech/sdr2000/sdr2000.htm.

8.0 TREC SDR Track Conclusions and Future

The SDR Track has been an enormous success with regard to its primary goals of bringing the speech recognition and information retrieval research communities together to explore the feasibility of implementing and evaluating retrieval from spoken audio recordings. Certainly, we have shown that the technology can be implemented and evaluated for TREC known item and ad hoc tasks. We've also found that it can be implemented and evaluated for reasonably large audio collections and for conditions where story boundaries are unknown. In fact, progress has occurred so quickly, that one might conclude that SDR is a solved problem. However, there is still much useful non-lexical information to be harnessed from the audio signal. Further, while we have explored traditional text retrieval modalities using automatically transcribed speech, we haven't yet tackled such challenging problems as question answering or spoken queries in which the mis-recognition of a single word could cause catastrophic failure of the technology. In our traditional SDR task, the redundancy of words in the collection has protected us from truly facing these issues. Finally, there are still many more issues to explore and conquer with regard to the more general problem of multi-media information retrieval.

There has been much discussion regarding the future of the TREC SDR Track and several suggestions for future evaluations revolving around an audio-only domain have been circulated including passage retrieval, multi-lingual or cross-lingual SDR, SDR with question answering, interactive SDR, to name a few. However, most of these problems are already being tackled on a text-only basis within TREC and, with the possible exception of question answering, the additional information to be learned from them for audio collections might be somewhat limited. We now have a fairly good idea of the kinds of problems that ASR introduces for text retrieval and we can most likely model the behavior of other text retrieval domains using ASR without running full-blown evaluations.

It seems to us that the next challenge is, rather, a broadening to a true multi-media information retrieval (MMIR) domain which will require not only text retrieval and speech recognition, but video and still image processing as well.[4] Further, these multi-media sources will come in many different forms which will need to be integrated and threaded. Such threading will no doubt require natural language processing and knowledge engineering. This is an enormous problem and will require collaboration among many different technology communities. For SDR, we brought together two research communities. MMIR will require the involvement of many more. Taken at once, this task seems virtually impossible. So, it will make sense to break it down into its constituent components or component combinations that can be incrementally integrated. Accordingly, we believe that several binary or ternary technology development and evaluation projects should be undertaken to explore the more tractable lower-level challenges before we undertake full MMIR. With this approach, core signal processing technologies such as speech recognition, speaker identification, face and object identification, scene tracking, etc. can be incrementally integrated with higher-level information processing technologies. Eventually, the capability to create robust multi-media information system technologies will emerge.

For next year, NIST is interested in creating a retrieval track that would begin to explore the information contained in the video signal. If a video corpus including audio is used, we can also begin to explore the integration of speech recognition and video processing into retrieval applications.

[4] Actually, we've only scratched the surface of audio processing with speech recognition, since a great deal more information than words are encoded in the audio signal.

These new domains and integrated technologies will, of course, require the development of new evaluation methods, formats, and tools. This is perhaps one of the greatest challenges to overcome in developing a new technology research task. For each of the research tasks that NIST has created evaluation programs for, there has been significant and sometimes lengthy discussion and debate regarding the development of metrics and scoring protocols. Metrics which are taken for granted today, such as mean average precision and word error rate, were once hotbeds of discussion. Further, we will need to build not only component technology measures, but end-to-end system measures as multi-media systems technologies take shape. The possibilities are quite exciting, but there is much work to be done.

Acknowledgements

NIST work in the TREC SDR tracks was sponsored in part by the Defense Advanced Research Projects Agency (DARPA).

The authors would like to thank Karen Spärck Jones at the University of Cambridge for her guidance in the development of the SDR Track. We'd like to thank Donna Harman and David Pallett at NIST for their support for the SDR track and Vince Stanford at NIST for his help in implementing the baseline speech recognition systems. We'd like to thank Sue Johnson at the University of Cambridge for her help in refining the test specifications and evaluation protocols. We'd like to thank IBM for their contribution of the baseline speech recognizer transcripts for TREC-6 SDR, Carnegie Mellon University for their contribution of the SPHINX-III recognizer for TREC-7 SDR, and a special thanks to GTE/BBN for the contribution and support of their LINUX-based BYBLOS Rough 'N Ready fast recognizer for use in TREC-8 SDR. Finally, we'd like all the TREC SDR participants without whose participation this track would not have been such a success.

Disclaimer

Bibliographical References

BEOWULF Project, NASA Center of Excellence in Space Data and Information Sciences, http://cesdis.gsfc.nasa.gov/linux/beowulf/, reviewed in 1997.

Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S, TDT-2 Text and Speech Corpus, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

Fiscus, J.G., Doddington, G., Garofolo, J.S., *NIST's 1998 Topic Detection and Tracking Evaluation*, Proc. 1999 DARPA Broadcast News Workshop, February 1999.

Fisher, re: investigation of TDT-2 transcription error rates, personal conversation, 1999.

Garofolo, J., Fiscus, J., and Fisher, W., *Design and preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora*, Proc. DARPA Speech Recognition Workshop, February 1997.

Garofolo, J., Voorhees, E., Stanford, V., and Spärck Jones, K., *TREC-6 1997 Spoken Document Retrieval Track Overview and Results*, Proc. TREC-6, 1997 and 1998 DARPA Speech Recognition Workshop, February 1998.

Garofolo, J. S., Voorhees, E. M., Auzanne, C.G.P. , Stanford, V.M., Lund, B.A., *1998 TREC-7 Spoken Document Retrieval Task Overview and Results*, Proc. TREC-7, Nov. 1998.

Garofolo, John S., Auzanne, Cedric G. P., Voorhees, Ellen M., *1999 Trec-8 Spoken Document Retrieval Track Overview and Results*, Proc. TREC-8, Nov. 1999. [Due to time constraints, the referenced paper was not created. Instead, this RIAO 2000 paper was also used as the TREC-8 SDR overview in the TREC-8 Proceedings]

Graff, D., Wu, Z., MacIntyre, R., and Liberman, M., *The 1996 Broadcast News Speech and Language-Model Corpus*, Proc. DARPA Speech Recognition Workshop, February 1997.

Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K., and Woodland, P.C., *The Cambridge University Spoken Document Retrieval System*, Proc ICASSP '99, Vol. 1, pp 49-52, March 1999)

Kantor, P., and Voorhees, E.M., *The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text*, Information Retrieval, In press – 2000.

Kubala, F., Colbath, S., Liu, D., Srivastava, A., Makhoul, J. *Integrated technologies for indexing spoken language*, Communications of the ACM, Volume 43, page 48, Feb. 2000.

Miller, D., Schwartz, R., Weischedel, R., Stone, R., *Named Entity Extraction from Broadcast News*, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

Pallett, D., Fiscus, J., and Przybocki, M., *1996 Preliminary Broadcast News Benchmark Tests*, Proc. DARPA Speech Recognition Workshop, February 1997.

Pallett, D.S., Fiscus, J.G., Martin, A., Przybocki, M.A., *1997 Broadcast News Benchmark Test Results: English and Non-English*, Proc. DARPA Broadcast News Transcription and Understanding Workshop, February 1998.

Pallett, D.S., Fiscus, J.G., Garofolo, J.S., Martin, A., Przybocki, M., *1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures*, Proc. DARPA Broadcast News Workshop, February 1999.

Przybocki, M.A., Fiscus, J.G., Garofolo, J.S., Pallett, D.S., *1998 Hub-4 Information Extraction Evaluation*, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

Singhal, A., Pereira, F., *Document Expansion for Speech Retrieval*, Proc. SIGIR '99, 1999.

Voorhees, E., Garofolo, J., and Spärck Jones, K., *The TREC-6 Spoken Document Retrieval Track*, Proc. DARPA Speech Recognition Workshop, February 1997.

Voorhees, E., Garofolo, J., and Spärck Jones, K., *The TREC-6 Spoken Document Retrieval Track*, TREC-6 Notebook, Nov. 1997.

Voorhees, E.M., Harman, D., *Overview of the Seventh Text REtrieval Conference (TREC-7)*, Proc. TREC-7, November 1998.

Voorhees, E.M., Harman, D., *Overview of the Sixth Text REtrieval Conference (TREC-6)*, Information Processing and Management, Vol. 36, No. 1, pp 3-35, January 2000.

TREC-6 SDR Participant Publications (http://trec.nist.gov/pubs/trec6/t6_proceedings.html)

Abberley, D., Renals, S., *The THISL Spoken Document Retrieval System*, University of Sheffield, UK, G. Cook, T. Robinson, Proc. TREC-6, Nov. 1997.

Allan, J., Callan, J., Croft, W.B., Ballesteros, L, Byrd, D., Swan, R., Xu, J., *INQUERY Does Battle With TREC-6*, Proc. TREC-6, Nov. 1997.

Crestani, F., Sanderson, M., Theophylactou, M., Lalmas, M., *Short Queries, Natural Language and Spoken Document Retrieval: Experiments at Glasgow University*, Proc. TREC-6, Nov. 1997.

Fuller, M., Kaszkiel, M., Ng, C.L., Vines, P., Wilkinson, R., Zobel, J. *MDS TREC6 Report*, Proc. TREC-6, Nov. 1997.

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., Schäuble, P., *ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval*, Proc. TREC-6, Nov. 1997.

Oard, D.W., Hackett, P., *Document Translation for Cross-Language Text Retrieval at the University of Maryland*, Proc. TREC-6, Nov. 1997.

Siegler, M.A., Slattery, S.T., Seymore, K., Jones, R.E., Hauptmann, A.G., Witbrock, M.J., *Experiments in Spoken Document Retrieval at CMU,* Proc. TREC-6, Nov. 1997.

Singhal, A., Choi, J., Hindle, D., Pereira, *F., AT&T at TREC-6: SDR Track*, Proc. TREC-6, Nov. 1997.

Smeaton, A.F., Quinn. G., Kelledy, F., *Ad hoc Retrieval Using Thresholds, WSTs for French Mono-lingual Retrieval, Document-at-a-Glance for High Precision and Triphone Windows for Spoken Documents*, Proc. TREC-6, Nov. 1997.

Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., Spärck Jones, K., *Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR,* Proc. TREC-6, Nov. 1997.

TREC-7 SDR Participant Publications (http://trec.nist.gov/pubs/trec7/t7_proceedings.html)

Abberley, D., Renals, S., Cook, G., Robinson, T., *Retrieval Of Broadcast News Documents With the THISL System* , Proc. TREC-7, Nov. 1998.

Allan, J, Callan, J., Sanderson, Xu, J., *INQUERY and TREC-7*, Proc. TREC-7, Nov. 1998.

Dharanipragada, S., Franz, M., Roukos, S., *Audio-Indexing For Broadcast News* (reference to TREC-6 SDR), Proc. TREC-7, Nov. 1998.

Ekkelenkamp, R., Kraaij, W., van Leeuwen, D., *TNO TREC7 site report: SDR and filtering*, Proc. TREC-7, Nov. 1998.

Fuller,. M., Kaszkiel, M., Ng, C., Wu, M., Zobel, J., Kim, D., Robertson, J., Wilkinson, R., *TREC 7 Ad Hoc, Speech, and Interactive tracks at MDS/CSIRO*, Proc. TREC-7, Nov. 1998.

Henderson, G.D., Schone, P., Crystal, T.H., *Text Retrieval via Semantic Forests: TREC7*, Proc. TREC-7, Nov. 1998.

Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K., Woodland, P.C., *Spoken Document Retrieval for TREC-7*, Proc. TREC-7, Nov. 1998.

Nowell, P., *Experiments in Spoken Document Retrieval at DERA-SRU*, Proc. TREC-7, Nov. 1998.

Oard, D.W., *TREC-7 Experiments at the University of Maryland*, Proc. TREC-7, Nov. 1998.

Siegler, M., Berger, A., Hauptmann, A., Witbrock, M., *Experiments in Spoken Document Retrieval at CMU*, Proc. TREC-7, Nov. 1998.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F., *AT&T at TREC7*, Proc. TREC-7, Nov. 1998.

TREC-8 SDR Participant Publications (http://trec.nist.gov/pubs/trec8/t8_proceedings.html)

Abberley, D., Ellis, D., Renals, S., Robinson, T., *The THISL SDR System At TREC-8*, Proc. TREC-8, Nov. 1999.

Allan, J., Callan, J., Feng, F-F., Malin, D., *INQUERY and TREC-8*, Proc. TREC-8, Nov. 1999.

Franz, M., McCarley, J.S., Ward, R.T., *Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM*, Proc. TREC-8, Nov. 1999.

Fuller, M., Kaszkiel, M., Kimberley, S., Ng, C., Wilkinson, R., Wu, M., Zobel, J., *The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC 8*, Proc. TREC-8, Nov. 1999.

Gauvain, J-L., de Kercadio, Y., Lamel, L., Adda, G., *The LIMSI SDR System for TREC-8*, Proc. TREC-8, Nov. 1999.

Han, B., Nagarajan, R., Srihari, R., Srikanth, M., *TREC-8 Experiments at SUNY Buffalo*, Proc. TREC-8, Nov. 1999.

Kraaij, W., Pohlmann, R., Hiemstra, D., *Twenty-One at TREC-8: using Language Technology for Information Retrieval*, Proc. TREC-8, Nov. 1999.

S.E. Johnson, P. Jourlin, K. Spark Jones, P.C. Woodland, *Spoken Document Retrieval for TREC-8 at Cambridge University*, Proc. TREC-8, Nov. 1999.

Siegler, M., Jin, R., Hauptmann, A., *CMU Spoken Document Retireval in TREC-8: Analysis of the role of Term Frequency TF*, Proc. TREC-8, Nov. 1999.

Singhal, A., Abney, S., Bacchiani, M., Collins, M., Hindle, D., Pereira, F., *AT&T at TREC-8*, Proc. TREC-8, Nov. 1999.