TREC-6 1997 Spoken Document Retrieval Track Overview and Results

John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford

National Institute of Standards and Technology (NIST)
Information Technology Laboratory
Building 225, Room A-216
Gaithersburg, MD 20899

Karen Sparck Jones

Cambridge University
Cambridge CB2 3QG, U.K.

ABSTRACT

This paper describes the 1997 TREC-6 Spoken Document Retrieval (SDR) Track which implemented a first evaluation of retrieval of broadcast news excerpts using a combination of automatic speech recognition and information retrieval technologies. The motivations behind the SDR Track and background regarding its development and implementation are discussed. The SDR evaluation collection and topics are described and summaries and analyses of the results of the track are presented. Finally, plans for future SDR tracks are described.

Since this was the first implementation of an evaluation of SDR, the evaluation itself as well as the evaluated technology should be considered experimental. The results of the first SDR Track were very encouraging and showed us that SDR could be successfully implemented and evaluated. However, the results of the SDR Track should be considered preliminary since the 50-hour spoken document collection used was very small for retrieval experiments (even though it was considered extremely large for speech recognition purposes.) Nonetheless, with thirteen groups participating in the TREC-6 SDR Track, a considerable amount of experience was gained in implementing and evaluating the SDR task. This experience will greatly benefit the next 1998 TREC-7 SDR Track.

1. MOTIVATION

Spoken Document Retrieval (SDR) involves the retrieval of excerpts from recordings of speech using a combination of automatic speech recognition and information retrieval techniques. In performing SDR, a speech recognition engine is applied to an audio input stream and generates a time-marked textual representation (transcription) of the speech. The transcription is then indexed and may be searched using an Information Retrieval engine. In traditional Information Retrieval, a topic (or query) results in a rank-ordered list of documents. In SDR, a topic results in a rank-ordered list of temporal pointers to relevant excerpts. In an operational SDR system, these excerpts could be topical sections of a recording of a conference or radio or television broadcasts. This technology when mature will permit users to search large collections of non-textual multi-media materials.

SDR was chosen as a TREC-6 (NIST Text REtrieval Conference 6) task for 1997 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 might work well enough now for usable SDR in some domains. SDR also provides a rich research domain in that it supports both development of large-scale near-real-time continuous speech recognition technologies and technologies for retrieval of spoken language. Further, SDR provides a venue for the development of synergy between the speech recognition and information retrieval communities to improve both technologies and create hybrids.

2. BACKGROUND

In November 1996 at the TREC-5 Workshop and later at the February 1997 DARPA Speech Recognition Workshop, NIST and Karen Sparck-Jones of Cambridge University held a discussion and a call for participation in a Spoken Document Retrieval (SDR) TREC Track for TREC-6. An SDR track would focus research on solving problems inherent in the retrieval of documents created by speech recognition technologies and in the recognition of large quantities of speech. Furthermore, the evaluation component of the track would permit the benchmarking of progress in the retrieval of documents corrupted by recognition errors.

It was decided that the track would involve retrieval of radio and television broadcast news recordings collected by the Linguistic Data Consortium (LDC) in 1996. The LDC Broadcast News (BN) corpus had been collected to support the DARPA-sponsored *Hub-4* continuous speech recognition project and was fully transcribed and annotated with story boundaries and could be adapted at little cost to the SDR task.[1] Both the CSR and IR communities expressed interest in the proposed project, so NIST and Sparck-Jones developed an evaluation plan to implement an initial SDR evaluation during the summer of 1997 to be reported at the November 1997 TREC-6 Workshop.

3. SDR EVALUATION PLAN

Initial discussion involved the nature of the retrieval task to be used in the SDR Track. It was acknowledged that because the amount of available Hub-4 Broadcast News corpora was limited and because this was to be a relatively low-overhead task, a full Ad-hoc-style TREC task was impractical. So, instead, a *known-item* task, which simulates a user seeking a particular, half-remembered document in a collection, was chosen. The goal in a known-item retrieval task is to generate a single correct document for each topic rather than a set of relevant documents as in an ad-hoc task. This approach simplified the selection of topics and eliminated the need for expensive relevance assessments. A known-item retrieval task had been successfully implemented in the similarly-designed TREC-5 *OCR Confusion Track*. [2]

The TREC-6 1997 SDR Evaluation Plan can be found at

http://www.nist.gov/speech/sdr97.txt

3.1 Evaluation Modes

The focus of the initial SDR evaluation was to encourage broad participation from both the Speech Recognition and Information Retrieval Communities. Therefore, the evaluation plan was designed to allow relatively easy entry for members of both communities. Speech recognition and retrieval experts were encouraged to team up to create hybrid SDR systems. The SDR Track included three retrieval conditions which provided control experiments as well as allowing sites without access to speech recognition technology to participate: [3]

Reference (S1) (required) – Retrieval using the "perfect" human-transcribed reference transcriptions of the Broadcast News recordings. This condition provided a control for retrieval.

Baseline (B1) (required) – Retrieval using the IBM-provided speech-recognition-system-generated 1-best transcriptions of the Broadcast News recordings. This condition provided a control for recognition and permitted sites without access to recognition technology to participate.

Full SDR (R1) (optional) – Retrieval using the Broadcast News recordings. This condition required both speech recognition and retrieval (which could be implemented by different sites).

Participants in the Full SDR condition with 1-best word-based recognizers were encouraged to submit the output of their recognition systems to be informally scored by NIST in evaluating the effect of recognition error rates on performance.

For purposes of simplifying the implementation and evaluation process, the hand-annotated temporal story boundaries were given in all conditions. Figure 1. Shows the SDR process for the TREC-6 task.

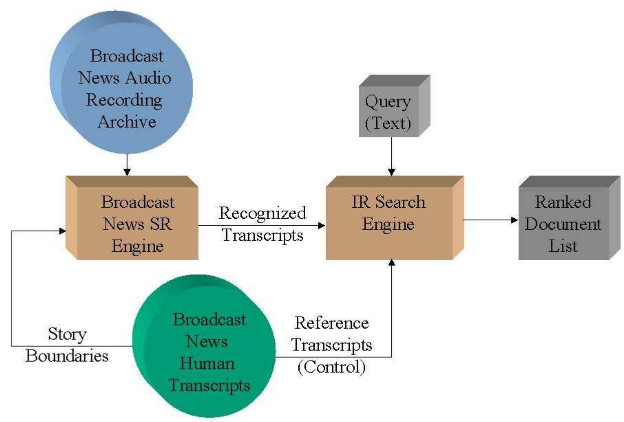

Figure 1. TREC-6 SDR Process

3.2 Test Corpora

The LDC Broadcast News corpus was chosen for the SDR task since it contained news data from several radio and television sources and was fully transcribed and pre-segmented by story.[1] To adapt the BN corpus to the SDR task, Story ID tags were added to uniquely identify each annotated story.

The existing 100 hours of broadcast news (which was originally collected by the LDC to provide training material for DARPA Hub-4 speech recognition systems) was divided into equal training and test sets. The 50-hour subset which was used for Hub-4 training in 1996 was chosen for SDR training and the newly-transcribed 50-hour subset was chosen as the test set for the SDR track.

This facilitated speech recognition site participation since sites with 1996 Hub-4 systems could apply them directly to the SDR task. [4] (Note that the two sets overlap temporally.)

An index was developed for the 50-hour test set to exclude commercials, sports summaries, weather reports, and untranscribed stories from the test. The baseline recognizer also had bad output for some sections of a few recordings. So that the Baseline test results could be directly compared to those for the Reference and Full SDR conditions, these stories were also removed from the test index.

The final filtered test set contained 1,451 stories with about 400,000 words. About 1/3 of the stories in the test set were labeled as "filler" – non-topical sections of the broadcasts. Because of the small size of the collection for retrieval testing, these were not removed from the test set. The mean length in words for the stories in the test set was 276 words. The histogram in Figure 2 shows the distribution of the length of the stories in the test set.

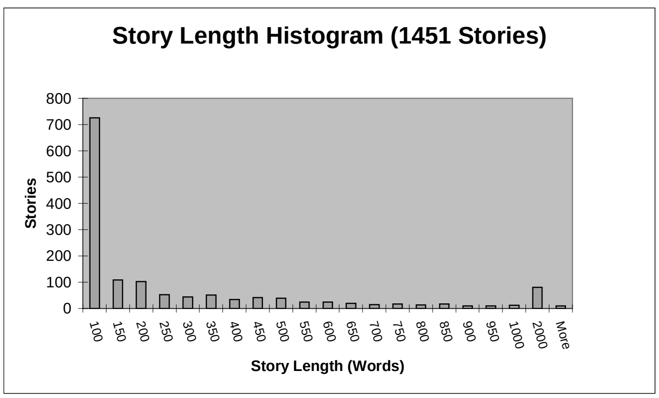

Figure 2. Test Set Story Length Histogram

Note that about half of the stories contain less than 100 words and a few stories contain 2000 or more words.

The recorded waveform material for the full SDR retrieval mode was made available to the participants in February, 1997. The human-created reference transcripts for the test collection and indices which specified the 1,451 usable stories were released in June. The test topics and baseline recognizer transcripts were released in the beginning of July and results were due at NIST in early September. The results of the SDR track were reported at TREC-6 in November 1997 and at the DARPA Broadcast News Transcription and Understanding Workshop in February 1998.

3.3 SDR Topics

As indicated in section 2.1, a known item retrieval task was selected for the SDR Track. Fifty known item topics, each intended to retrieve a single spoken document, were selected at NIST – half to exercise the retrieval challenges of the task and half to exercise the speech recognition challenges of the task.

To this end, 25 topics were selected by the NIST Spoken Natural Language Processing and Information Retrieval Group to pose various challenges to the retrieval systems. This topic subset is referred to later in this paper as **"Difficult Topics"**. An example "difficult" topic (SDR3) was: *What is the difference between the old style classic cinemas and the new styles of cinema we have today?* This topic targeted a story (CNN Headline News, June 7 1996, story 28) which did not contain the word, "cinema". Instead, the document contained several instances of the synonym, "theater". So, this query required systems to use synonymy to retrieve the target story.

The remaining 25 topics were selected by the NIST Spoken Natural Language Processing Group to cover the spectrum of the speech recognition challenges of the task. and divided into two subcategories to emphasize two Hub-4 speech recognition conditions:

> **"Easy" Speech (Hub-4 F0):** High fidelity recording, non-spontaneous speech, native speaker of American English, quiet conditions. An example "F0" topic (SDR1) was: *Does the Olympic torch ever travel by motorcycle?* This topic targeted a story with a scripted reading by a news anchor under low noise/high bandwidth conditions. This story (NPR All Things Considered, June 18 1996, story 9) contained none of the Hub-4-categorized phenomena thought to cause recognition difficulty.
>
> **"Difficult" Speech (Hub-4 FX):** Combinations of speech recognition degrading conditions such as low fidelity channel, spontaneous speech, non-native speaker, noisy conditions. An example "FX" topic (SDR33) was: *In what country do parents fear that the devil is going to come and take their children?* This topic targeted a story (with "medium" fidelity, "high" background noise, and several areas of non-English-speaking speakers with interpreters. Ninety two percent of the material in the story (CNN Headline News, June 7 1996, story 12) included 2 or more Hub-4-categorized phenomena thought to cause recognition difficulty.

Each of these topics was selected to target at story which contained primarily either Easy (F0) or Difficult (FX) categorized speech. These topic subsets are referred to later in this paper as "Easy Speech" and "Difficult Speech".

One of the 50 topics was removed from the test because it retrieved a story which was in an errorful set of output produced by the baseline recognizer. So, the retrieval for 49 topics was scored and tabulated in the SDR Appendix of the TREC-6 notebook. It was also discovered that two topics properly retrieved multiple stories from the test set. Since this was a known-item task, for simplicity, these were removed in the analyses provided in this paper. Therefore, the results presented in this paper are based on only 47 topics.

4. EVALUATION RESULTS

In all, 13 sites (or site combinations) participated in the SDR Track. Nine of these performed the speech recognition portion as well as retrieval portions of the task and implemented the Reference, Baseline, and Full SDR retrieval conditions:

- AT&T
- Carnegie Mellon University Informedia Group
- Claritech (with CMU Speech Recognition)
- ETH Zurich
- Glasgow (with Sheffield University Speech Recognition)
- IBM
- Royal Melbourne Institute of Technology
- Sheffield University
- University of Massachusetts (with Dragon Systems Speech Recognition)

The remaining 4 sites implemented only the Reference and Baseline retrieval conditions:

- City University of London
- Dublin City University
- University of Maryland
- NSA

4.1 Speech Recognition Component Performance

The primary purpose of the SDR Track was to evaluate the retrieval of spoken documents and not speech recognition. To this end, there was no formal evaluation of the speech recognition component of the Full SDR systems. However, if sites used 1-best word recognition to produce retrieval transcripts for Full SDR, they were encouraged to submit these so that NIST could exam the relationship between word error rate and retrieval performance. Of the eight participating Full SDR sites, 4 submitted recognition output to NIST for scoring (one of these was the IBM-contributed baseline recognizer.) Other Full SDR sites either used another site's recognizer, used an alternative recognition technique such as phone recognition, or choose not to share their recognition results. Figure 3 shows a histogram of the story Word Error for the IBM system.

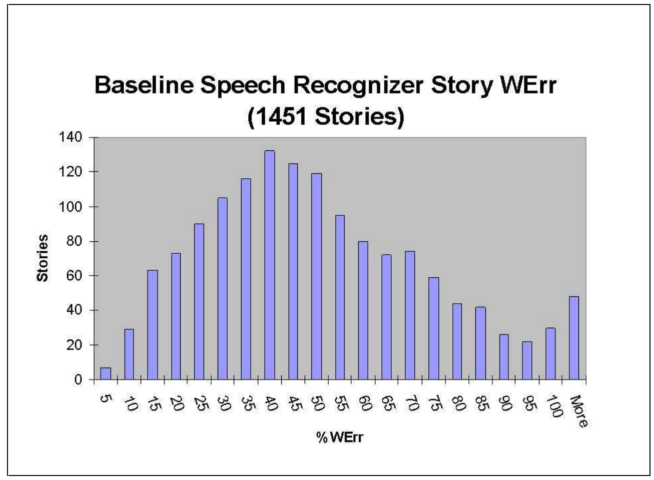

Figure 3. Story Word Error Rate for the IBM Recognizer

The story Word Error Rate mode for the baseline recognizer was approximately 40% while the mean was substantially higher at 50.0% because of the long right tail in the distribution. The mean story Word Error Rate for the other recognizers fell between approximately 35% and 40%.

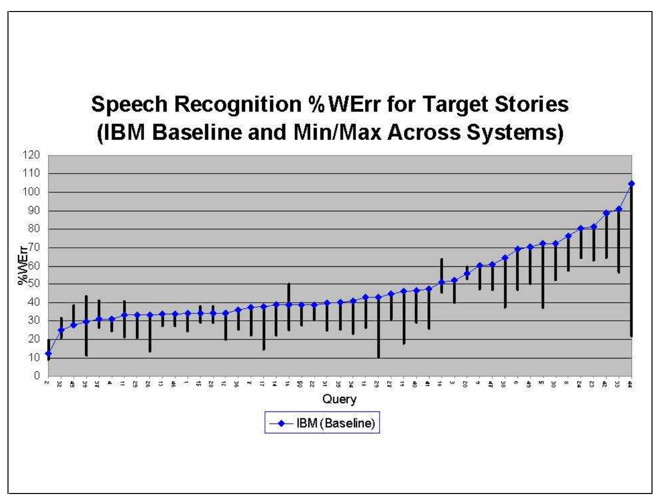

Figure 4. Sorted Target Story Baseline Recognizer Word Error Rate with Min/Max for other Recognizers

The Word Error Rate for each of the 47 target stories was sorted by increasing error for the Baseline recognizer and plotted along with the min and max Word Error Rates for the other submitted recognizers in Figure 4. Note that the plot for the Baseline recognizer shows a fairly good distribution of error rates across the target story subset.

The recognizer Word Error Rates were determined using procedures and software (sclite) similar to those used in the NIST/DARPA 1996 Hub-4 Broadcast News Continuous Speech Recognition Benchmark Tests. Once scored, the error rates were tabulated by story rather than speaker, segment, or focus condition as in Hub-4.[5] However, the SDR reference transcriptions were not checked and corrected as in Hub-4 and the 1996 Hub-4 orthographic mapping file for lexical normalization was employed which provided only minimal coverage of the SDR test set.[4] Therefore, these SDR Word Error Rates **cannot** be directly compared to those for Hub-4 systems. However, they do provide a point of reference for measuring the relative difficulty of retrieval of stories with respect to recognition accuracy.

4.2 Retrieval Results

Test participants were required to submit a relevance-rank-ordered list of the ID's of the top 1000 stories they retrieved for each topic. But, since the SDR Track employed a known-item task, the results of the retrieval for a topic were considered to be correct only if the target document for the topic appeared at rank 1.

In evaluating retrieval performance, we investigated the measures used in the TREC-5 Confusion Track[2]:

> **Mean Rank When Found** – mean rank at which the target story was found averaged across all topics that retrieved the target story in the top 1000 documents.
>
> **Mean Reciprocal Rank –** mean of the reciprocal of the rank at which the target story was found over all the topics using 0 as the reciprocal for topics that did not retrieve the story.

Another measure, a plot of the number of topics that retrieve the target document by a certain rank was suggested by ETH.

These measures as well as the rank at which each topic was found were reported in the SDR Appendix of the TREC-6 Notebook.[6]

The results for all three test conditions (Reference, Baseline, and Full SDR) were surprisingly good for an initial evaluation of retrieval of spoken language transcripts. Retrieval rates were very high for the human-transcribed Reference data and most sites showed only small degradation in performance for Full SDR using their own recognition technology. There was generally more degradation using the Baseline recognizer transcripts due to its higher error rate and probably also due to a higher number of "out of vocabulary" (OOV) words. An exception was the Dublin system which showed slightly better performance for the Baseline than the Reference.

Since the results were very good, we decided to employ an additional measure, **Percent Retrieved at Rank 1** across systems and test conditions, which is shown in Figure 5.

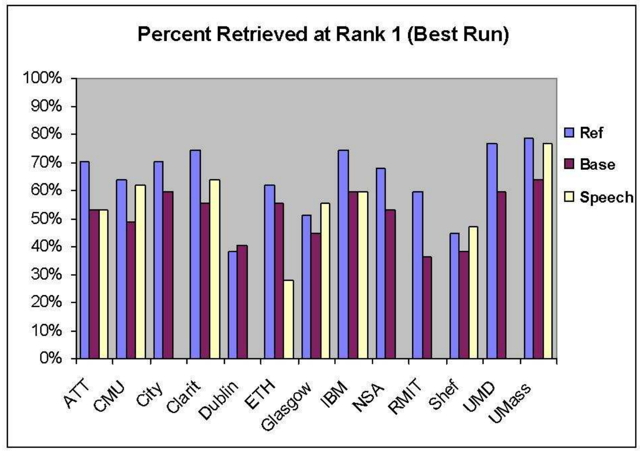

Figure 5. Retrieval rate at rank 1 for all systems and modes (best run)

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). The UMass system yielded a retrieval rate of 78.7% for the Reference mode, 63.8% for the Baseline mode, and 76.6% for Full SDR. Note that the UMass Reference and Full SDR results differed by only one topic.

A comparable graph for the Mean Reciprocal Rank is given in Figure 6.

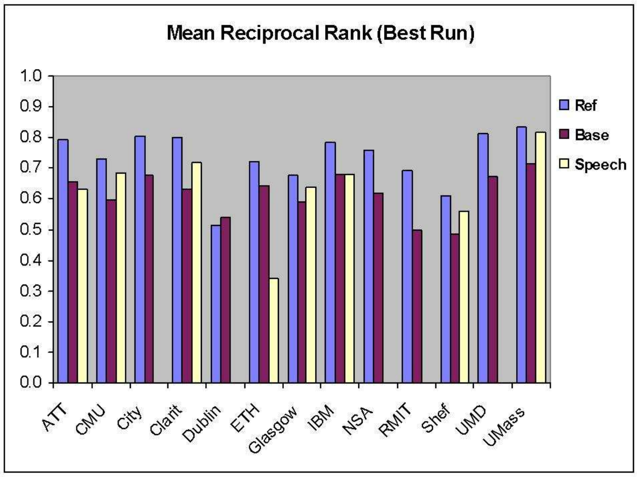

Figure 6. Mean Reciprocal Rank for all systems and modes (best run)

For this evaluation, the Percent Retrieved at Rank 1 and Mean Reciprocal Rank metrics did not show significantly different relative system ranks or trends. It is interesting to note, however, that for the Percent Retrieved at Rank 1 measure only, the Glasgow and Sheffield systems performed more poorly on the Reference condition than on Full SDR most likely due to a bug in processing the Reference transcripts.

Although there is disagreement between the two measures above (Percent Retrieved at Rank 1 and Mean Reciprocal Rank) for the relative ranking of the performance of the retrieval modes for Sheffield and Glasgow, a regression test of the Percent Retrieved at Rank 1 versus the Mean Reciprocal Rank (Figure 7) shows that the two measures were not significantly different for this evaluation.

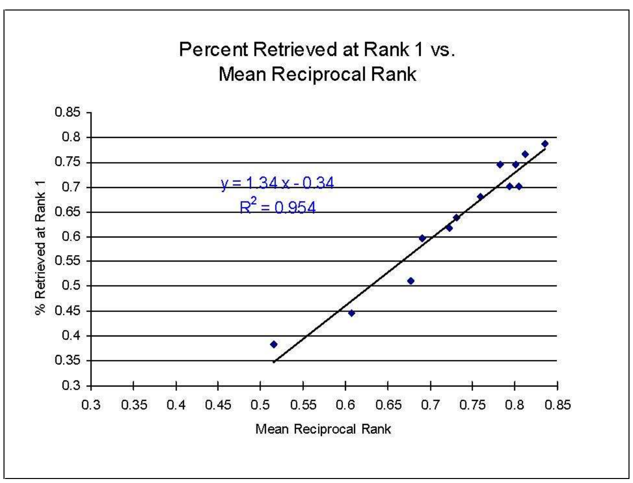

Figure 7. Regression Plot of Percent Retrieved at Rank 1 vs. Mean Reciprocal Rank

An examination of Percent Retrieved at Rank 1 averaged across systems for each of the topic subset (Figure 8) shows that The "Easy to Recognize" (F0) topic/story set yielded the best performance for all 3 evaluation modes (Ref, Base, and Full SDR) and the "Difficult to Recognize" (FX) topic/story set yielded significantly degraded performance. However, the "Difficult Query" subset yielded even greater performance degradation.

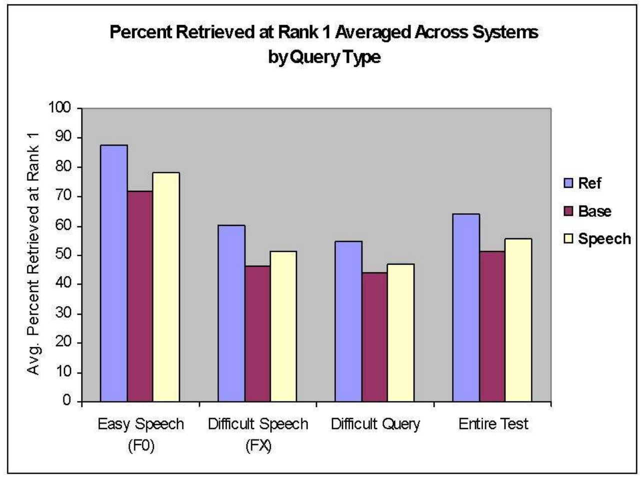

Figure 8. Percent Retrieved at Rank 1 averaged across systems by topic subset

It is interesting to note that in general, the systems had difficulty with the "Difficult Speech" topics for the Reference retrieval mode (in which the target story texts were not degraded by recognition errors) as well as the Baseline and Full SDR modes (which contained recognition errors.) This may indicate a relationship between language characteristics that degrade recognition and factors that make it difficult to retrieve a spoken document. However, this hypothesis is confounded with the effect topic difficulty had on retrieval. Figure 8 also shows that the topic difficulty had a much greater effect on retrieval performance than retrieval mode (Reference, Baseline, Full SDR. So, it is clear for future SDR evaluations that if the relationship between recognition and retrieval is to be explored, topic difficulty factors will need to be controlled or at least measured.

In order to look at recognition-related retrieval error trends, we overlayed the sorted Baseline recognizer story Word Error Rate from Figure 4 over the rank at which each story was retrieved (mean, min, and max) across systems for each retrieval mode. This is shown in Figure 9 for the Baseline retrieval condition and in Figure 10 for the Full SDR retrieval condition.

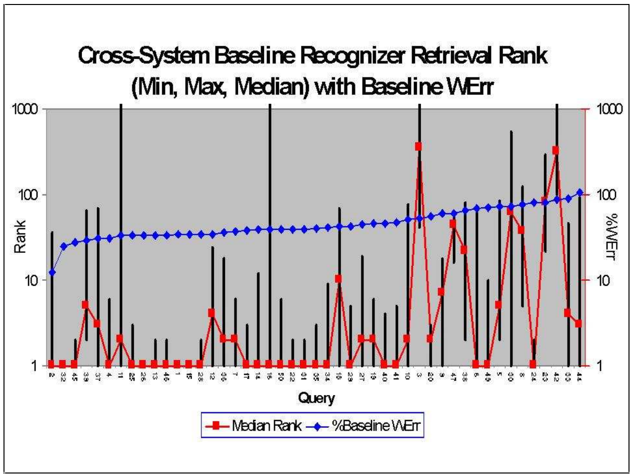

Figure 9. Baseline retrieval mode target story median, min, and max retrieval rank averaged across systems sorted by Baseline Recognizer story Word Error

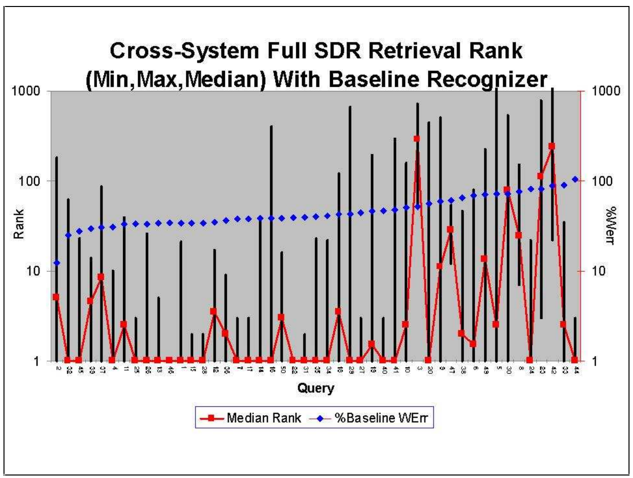

Figure 10. Full SDR retrieval mode target story median, min, and max retrieval rank averaged across systems sorted by Baseline Recognizer story Word Error

Note that the mean ranks appear to indicate a trend toward increasing retrieval error as the target story recognition error rate increases. However, the same plot for the Reference retrieval condition shown in Figure 11 (which did not suffer from recognition errors) shows a surprisingly similar trend. It appears that difficult-to-recognize stories are also difficult to retrieve -- even if the "perfect" transcribed version of the stories is used for retrieval. This may indicate that there is an indirect relationship between recognition difficulty and retrieval difficulty at the lexical level. One hypothesis is that the complexity of the language itself in these difficult stories is greater. They may also contain fewer key content-bearing words.

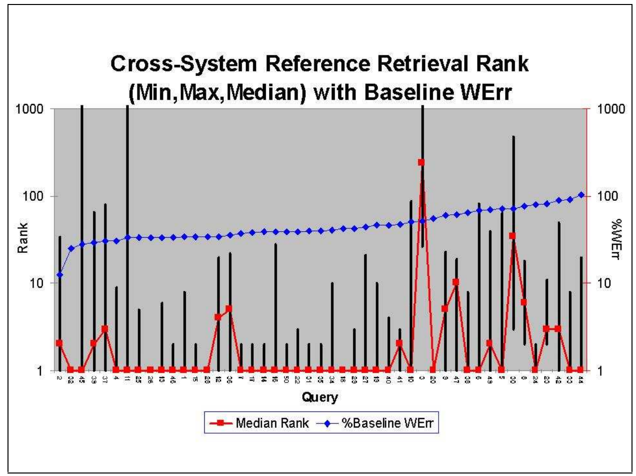

Figure 11. Reference retrieval mode target story median, min, and max retrieval rank averaged across systems sorted by Baseline Recognizer story Word Error

An interesting exception is found in the results for SDR18. The SDR18 topic was: *Has D.N.A. evidence been used in the Unabomber case?* All thirteen systems were able to correctly retrieve the target story for this topic in the Reference Retrieval condition. But, only 2 of the 13 systems were able to correctly retrieve the story in the Baseline Retrieval condition and only 3 of the 8 systems implementing Full SDR were able to retrieve the document. Upon examining the recognized transcriptions for this story, we find that the key content word, "Unabomber" is never correctly recognized and most systems also had difficulty with a secondary key word, "evidence". These words were most likely "out of vocabulary" for the recognition systems. The retrieval systems failed to retrieve the story since these key content-bearing words were lost. This kind of problem had only a small impact on retrieval using this very small test collection. However, the OOV problem could have a much more substantial impact on retrieval using realistically large spoken document collections.

Next year, measures of *content word* story recognition may be employed which would provide a better picture of the relationship between recognition accuracy and retrieval performance.

4.3 Statistical Analyses

ANOVA statistical significance tests were also implemented to measure the relative importance of each of the SDR component technologies (recognition and retrieval) in contributing to SDR retrieval performance. When comparing variance from differences in sites and retrieval modes (Reference Baseline, Full SDR), we found that the experimental design was adequate to highlight differences across systems in the Reference and Baseline modes, but not for Full SDR. When comparing variance for the site and the Reference and Baseline retrieval modes, ANOVA evaluation showed that 66.5% of the variance was attributable to the site, 26.3% to the retrieval mode, and only 7.2% was unexplained. (Similar results were also observed if only the subset of sites who performed Full SDR were evaluated.) Note that this result indicates that the retrieval method used was almost three times more important than the transcript (or recognizer) used in differentiating systems.

However, when the variance for the site and Reference, Baseline, and Full SDR retrieval modes was compared, ANOVA evaluation showed that 57.1% of the variance was attributable to the site, 18.9% to the retrieval mode, and **24.0%** to unexplained factors -- thus indicating that effects from the interaction of CSR and IR components were confounded in the results.

This problem would be eliminated only if all IR components were combined with all CSR components and evaluated. An exhaustive cross-component comparison is impractical, at least for the near future. But, CSR sites who produce one-best recognized transcripts will be encouraged to share these with the other participants so that retrieval runs (which are relatively inexpensive) can be run with different recognizer transcripts. This should significantly reduce the problem with unexplained variance.

5. CONCLUSIONS

The first evaluation of SDR technology showed that relatively good known-item retrieval could be achieved for a small collection of broadcast news spoken documents. It also showed that existing speech recognition and information retrieval technologies could be effectively pipelined to perform the task and that spoken document retrieval as well as the underlying component technologies could be evaluated. The initial task, although small by IR standards, brought the IR and CSR communities together and initiated dialogue and collaboration.

During discussions at TREC-6, the SDR participants generally 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 and should be replaced with an ad-hoc-style relevance evaluation using pooled topics.

Since the test collection this year was far too small to simulate a real deployment of SDR technology, it is impossible to make sweeping judgements about the performance of SDR for real tasks or how well current approaches will scale. It is also, therefore, difficult to make conclusions regarding the relative importance of speech recognition and retrieval accuracy in overall retrieval performance or in the scalability of the technology.

For this test, it appeared that recognition accuracy was not nearly as an important factor as search performance in determining overall retrieval performance. However, it is highly possible that this relationship will not hold for realistically large spoken document collections. In any case, future SDR evaluations with much larger spoken document collections and relevance assessment should help to answer these questions.

6. FUTURE

To progress toward these goals, it is planned that the TREC-7 SDR Track will expand to include a 100-hour test collection and 25 Ad-Hoc-style topics to be developed by the NIST TREC assessors. Like in the TREC Ad-Hoc Track, The retrieved list of stories provided by the participating systems for each topic will be pooled and assessed for relevance by the assessors. Traditional Precision/Recall scoring will then be applied to the results.

For TREC-8, it is planned that the Broadcast News portion of the new TDT-2 Corpus will be used to provide a much larger and more realistic test collection – an order of magnitude larger than the TREC-7 SDR test collection. Current plans call for 1,000 hours/40,000 stories of Broadcast News by the end of 1998.

7. ACKNOWLEDGEMENTS

The authors would like to thank David Pallett of NIST for his contributions in development of the concept and design of the SDR Track. We'd also like to thank Jon Fiscus, Darrin Dimmick, and Carol Barnes at NIST for their assistance in tabulating the speech recognition and retrieval scores. Finally, we'd like to particularly thank Salim Roukos and Satya Dharanipragada of IBM for their contribution in providing the baseline recognizer transcriptions for the track.

NOTICE

REFERENCES

1. 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.

2. Kantor, P., and Voorhees, E.M., *The TREC-5 Confusion Track*, Proc. TREC-5, November 1996.

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

4. 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.

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

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