

Spoken Document Retrieval: 1998 Evaluation and Investigation of New Metrics

John S. Garofolo, Ellen M. Voorhees, Cedric G. P. Auzanne, Vincent M. Stanford

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

ABSTRACT

This paper describes the 1998 TREC-7 Spoken Document Retrieval (SDR) Track which implemented an 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. Alternative metrics for automatic speech recognition as applicable to retrieval applications are also explored. Finally, plans for future SDR tracks are described.

1. BACKGROUND

Spoken Document Retrieval (SDR) involves the search and 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.

The first community-wide evaluation SDR technology was implemented in 1997 for TREC-6. This pilot evaluation implemented a "known-item" task in which a particular relevant document was to be retrieved for each of a set of queries over a 50-hour collection of radio and television news broadcasts. Three retrieval conditions were implemented to examine the effect of recognition performance on retrieval performance:

1. *Reference* - retrieval using human-generated reference transcripts which for the purposes of this evaluation were considered to have "perfect" recognition.
2. *Baseline* - retrieval using IBM-contributed recognizer-generated transcripts with a 50% Word Error Rate. This provided both a common

recognition error condition and an entrée for sites who did not have access to a recognition system of their own.

3. *Speech* - retrieval using the recordings of the broadcasts themselves requiring both recognition and retrieval technologies.

Thirteen sites participated in the pilot evaluation, eight of which implemented the Speech retrieval condition using their own or a team site's speech recognition system. The pilot evaluation proved that an evaluation of SDR technology could be implemented and that existing technologies worked quite well for a known-item task on a small collection. The results were so good, that NIST chose to highlight the percent of target stories which were retrieved at rank one by the systems.

Using the Percent Retrieved at Rank 1 metric, the University of Massachusetts retrieval system yielded the best performance in all three conditions. The UMass system achieved a retrieval rate of 78.7% for the Reference Retrieval condition and 63.8% for the Baseline Retrieval condition. For the Full SDR condition, UMass using a Dragon-Systems-produced 1-best recognizer transcript with a 35% Word Error Rate, achieved a 76.6% retrieval rate.[1]

2. MOTIVATION

The 1998 SDR Track was designed to address the known inadequacies in the 1997 SDR Track (small corpus, known-item task) to provide a more realistically challenging retrieval task. For 1998, a nominally 100-hour broadcast news test set collected by the Linguistic Data Consortium (LDC)[1], was selected and a traditional TREC ad-hoc-style relevance task was chosen with topics and relevance assessments generated by human assessors. Two recognizer-produced transcript sets with different word error rates were provided by NIST as well as LDC human-generated reference transcripts. Also, for the first time, sites were encouraged to contribute their one-best recognizer-produced transcripts so that other sites could run retrieval on them. The improved test paradigm and alternative transcription

sets with a spectrum of recognition error rates permitted us to further examine the relationship between recognition errors and retrieval accuracy. The new cross-recognizer task also permitted us to explore the development of alternative metrics for automatic speech recognition technology which would address particular inadequacies of the technology with regard to its use in information retrieval applications. [2]

3. SDR EVALUATION PLAN

The complete evaluation plan for the 1998 TREC-7 Spoken Document Retrieval Track can be found at:

<http://www.nist.gov/speech/sdr98/sdr98.htm>

3.1 Evaluation Modes

The SDR Track included four retrieval conditions which provided component control experiments as well as allowing sites without access to speech recognition technology to participate:

Reference (R1) (required) – Retrieval using the “perfect” human-transcribed reference transcriptions of the Broadcast News recordings.

Baseline (B1/B2) (required) – Retrieval using two sets of speech-recognition-generated 1-best transcripts produced by NIST using the CMU SPHINX-III recognition system. B1 had a moderate (33.8%) word error rate and B2 a substantially higher (46.6%) word error rate.

Speech (S1/S2) (optional) – Retrieval using the Broadcast News recordings. This condition required both speech recognition and retrieval (which could be implemented by different sites).

Cross Recognizer (CR) (optional) - Retrieval using 1-best speech-recognizer-generated transcripts contributed by other sites. This condition provided a control for recognition as well as allowing us to evaluate retrieval using a variety of recognition systems with a range of error rates.

Speech recognition and retrieval experts were encouraged to team up to create pipelined or hybrid SDR systems. In addition, two participation levels were created to allow participation by retrieval sites which did not have access to a speech recognition system:

Quasi-SDR: Sites without access to speech recognition technology were permitted to run retrieval on only the baseline recognizer

transcripts and reference transcripts. (Retrieval conditions R1, B1, B2 only)

Full-SDR: Sites implemented both recognition and retrieval on the recorded news broadcasts as well as retrieval on the baseline recognizer transcripts and reference transcripts. (Retrieval conditions R1, B1, B2, and S1 minimally)

Participants in Full SDR with 1-best word-based recognizers were encouraged to submit their recognized transcripts to NIST for recognition vs retrieval evaluation to provide the material for the CR condition.

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]

A subset of 100 hours of the Broadcast News Corpus collected between June 1997 and January 1998 was chosen as the test corpus. In all, 87 hours was selected after filtering. It contained 2,866 stories with about 772,000 words.

The evaluation was run during the summer of 1998 and the results were reported at TREC-7 in November 1998 and at the DARPA Broadcast News Workshop in March 1999.

3.3 Baseline Recognizer Transcripts

CMU permitted NIST to use its SPHINX-III broadcast news recognition system to create a set of recognition-generated transcripts for the baseline retrieval (B1/B2) condition.

NIST used a 40 nodes cluster of PC-based Linux machines to implement 2 baseline recognition runs. The first (B1) system yielded a 33.8% word error rate and the second one (B2) a 46.6% word error rate.

3.4 SDR Topics

A team of NIST TREC assessors met in April 1998 to select 25 topics for the test collection using similar procedures to those used in other TREC ad-hoc tasks. The assessors were instructed to find topics with 7 or more relevant news stories each in the collection using the NIST PRISE search engine. Because of limitations in the collection, however, the assessors were able to develop only 23 such topics. The mean number of relevant stories for each topic is 16.96. The following are samples:

Find reports of fatal air crashes (Topic 62)

What economic developments have occurred in Hong Kong since its incorporation into the Chinese People's Republic (Topic 63)

4. EVALUATION RESULTS

In all, 11 sites (or recognition/retrieval teams) participated in the SDR Track. Eight of these sites performed the Full SDR task by implementing both the recognition and retrieval components (S1). These sites were also required to implement the R1, B1, and B2 control conditions.

Full SDR (recognition and retrieval - R1, B1, B2, S1):

- AT&T (ATT)
- CMU Group 1 (CMU1)
- Cambridge University, UK (CUHTK)
- DERA, UK (DERA)
- Royal Melbourne Institute of Technology, Australia (MDS)
- Sheffield University, UK (SHEF)
- TNO-TPD TU-Delft, Netherlands (TNO)
- University of MA - Dragon Systems (UMass)

The remaining 3 sites performed only the Quasi-SDR portion of the task.

Quasi-SDR (retrieval only - R1, B1, B2):

- CMU Group 2 (CMU2)
- NSA (NSA)
- University of MD (UMD)

4.1 Speech Recognition Component Performance

The primary purpose of the SDR Track was to evaluate the retrieval of spoken documents. However, if sites used 1-best word recognition to produce transcripts as input to their retrieval systems, they were encouraged to submit these for sharing in the Cross-Recognizer condition and so that NIST could examine the effect of recognition error rate and retrieval performance.

Of the 8 participating Full SDR sites, 5 submitted recognition output to NIST for scoring. Other Full SDR sites either used an alternative recognition technique such as phone recognition or lattices or chose not to share their recognition results.

The best recognition results were submitted from the Cambridge University HTK recognition system with a 24.6% test set word error rate and a 22.2% mean story word error rate.[4] Figure 1 shows a histogram of the

story word error rates for each of the submitted 1-best systems. This histogram gives a graphical profile of recognizer performance over the 2866 stories in the collection. A complete table of recognition scores for the submitted systems is given in Appendix A.

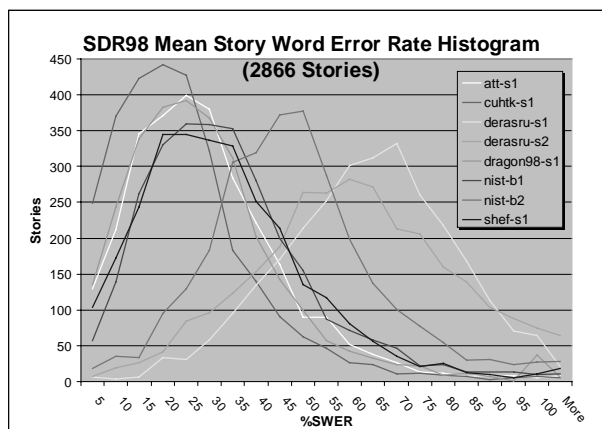


Figure 1. Story Word Error Rate Histograms for Submitted Recognized Transcripts

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. These results were then scored against the reference assessments created by the NIST assessors using the TREC_EVAL scoring software. As in other TREC tasks, the primary retrieval metric for the SDR evaluation was mean average precision (MAP) over all topics. Figure 2 shows the results for the R1, B1, B2, S1, and S2 retrieval conditions.

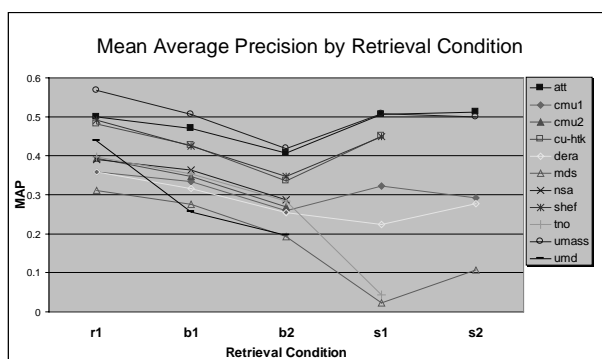


Figure 2 . Mean Average Precision for Required Retrieval Conditions

The graph shows that for all retrieval conditions except S2, the University of Massachusetts system achieved the best mean average precision. The UMass system achieved a MAP of .5668 for retrieval using the human reference transcripts (R1). The same system applied to

the moderate error baseline recognizer (B1) transcripts achieved a MAP of .5063, and for the high error baseline recognizer (B2) transcripts, a MAP of .4191. For the Speech input condition (S1) using their own team site's (Dragon Systems) recognizer at 29.5% word error rate, the UMass system achieved a MAP of .5075.[5] The AT&T system performed similarly for the S1 condition with a MAP of .5065. AT&T implemented a second recognition/retrieval system (S2) which achieved a MAP of .5120 - the highest results for input from a recognizer in this evaluation. It is interesting to note that the AT&T S1 and S2 results exceeded the results AT&T obtained (.4992 MAP) for the human reference transcripts (R1). AT&T attributes this to a new approach they implemented for document expansion using contemporaneous newswire texts. They applied the new document expansion approach to only their S1 and S2 runs and not on their other runs.[6] Appendix A gives a complete tabulation of the mean average precision scores for all of the systems and conditions.

In general, the results for this evaluation were quite good, with a seemingly linear but gentle decline in mean average precision for recognition transcripts with higher word error rate. The Cross-Recognizer retrieval results were used to further explore this apparent relationship.

4.3 Cross-Recognizer Retrieval Results and Alternative Recognition Metrics

The cross testing of several retrieval systems against several recognizer transcripts sets permitted us to examine retrieval performance over a wider variety of recognition data and we could begin to truly examine the relationship between recognition performance and retrieval performance. It also provided us with data to evaluate our recognition metrics for their suitability for retrieval and to experiment with new ones as well.

Four of the Full SDR sites: Cambridge University, DERA, RMIT/MDS, and Sheffield University implemented the Cross-Recognizer (CR) retrieval condition on the submitted recognized transcripts which covered a wide range of word error rates. See Appendix A for a tabulation of the results.

When we plot mean story word error rate against mean average precision for each of the 4 retrieval systems (Figure 3), we see a linear trend in mean average precision as average recognition word error increases. The R-squared statistic tells us that for these 4 retrieval systems over the 9 recognition points, there is a .8694 average correlation. This high correlation indicates that there is indeed a significant relationship between word error rate and retrieval accuracy. However, the retrieval results for the B2 recognizer are inexplicably low (far

below the results for the DERA recognizers with significantly higher word error rates). The plot also shows us that the performance for all of the retrieval systems remains relatively parallel with respect to recognizer.

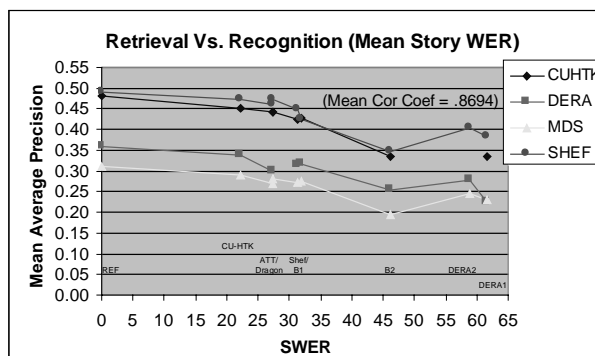


Figure 3. Cross-Recognizer Results: Mean Average Precision vs. Mean Story Word Error Rate

We believed, however, that we might achieve an even higher correlation if we employed a metric for speech recognition which emphasized the information-carrying words which are key for retrieval. Such a metric would be more predictive of retrieval performance and could be used to determine the suitability of a recognizer for use in a retrieval task. We evaluated 2 broad types of metrics:

Named-Entity-based: This metric would evaluate the error rate for named-entity words (people, locations, and organizations) as defined in the 1998 Hub-4 Information Extraction - Named Entity Evaluation.[7] The disadvantage of this metric is that it requires named-entity annotations in the reference transcripts. However, to our great fortune, GTE/BBN had annotated SDR reference transcriptions for use as named-entity training data.[8] We developed the following metric:

named entity word error rate (ne-wer): score only the named entities in the recognizer transcripts. To implement ne-wer, we used IE-Eval/REEP Named Entity scoring software [9] to align the annotated named entity words in the reference transcript with words in the recognizer transcripts. The alignments (with embedded named-entity tags) were then scored using the NIST SCLITE speech recognition scoring software. The embedded tags permitted us to score only named-entity words. So that we did not introduce entity tagger error into our metric, we ignored named entity words which might be inserted by the recognizer and evaluated only named entity words as annotated in the reference transcripts.

General IR-based: These metrics would use IR approaches themselves to process, filter, and weight the words in the recognizer transcripts to be scored. Such metrics might be useful in predicting retrieval performance based on recognition performance and could, therefore, be used to tune a recognizer for a retrieval task. We considered 3 such metrics:

stop-word-filtered word error rate (swf-wer): apply a stop-word list to the words in the reference and recognizer transcripts to remove stop (non-information-carrying) words. To implement this metric, we removed all occurrences of words in a 396-word stop word list from both the reference and recognizer transcripts. We then performed SCLITE word error rate scoring on the filtered transcripts.

Stemmed stop-word-filtered word error rate (sswf-wer): apply a stemmer to the results of the swf-wer filtering process above to remove word differences which are irrelevant to retrieval algorithms. To implement this metric, we applied an implementation of the Porter stemmer [10] to the stop-word-filtered reference and recognizer transcripts. We then performed SCLITE word error rate scoring on the filtered transcripts.

IR-weighted stemmed stop-word-filtered word error rate (IRW-WER). Apply an IR indexing algorithm to weight words prior to SCLITE word error rate scoring. We are currently examining IR algorithms for this application and have not yet implemented this metric.

The results of these alternative metrics as applied to the SDR recognizer are shown in Figure 4. Note that only the Named-Entity-based metrics appear to clearly change the relative ranking of the recognizer transcript sets. These metrics also show that the B2 recognition system was a poorer performer with regard to named entities than is evidenced by its word error rate. Our hypothesis is that the adjustment we made to the SPHINX pruning thresholds artificially reduced the likelihood of longer words being recognized - words which are more likely to be content-carrying named entities. The scores for each of the metrics as applied to each of the recognized transcripts are given in Appendix A.

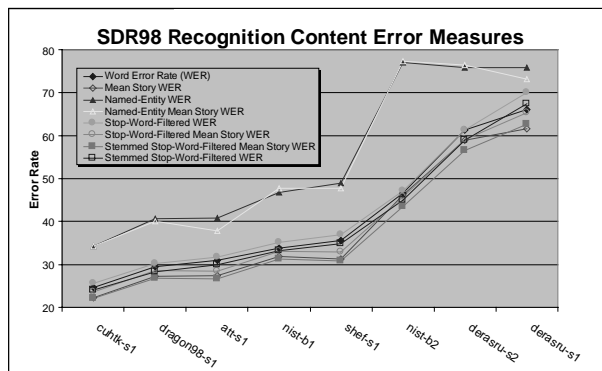


Figure 4 . Alternative Recognition Metrics

To quantify the efficacy of these metrics as predictive tools, we performed a R-squared analysis on the scores for the 4 retrieval systems versus the recognition metrics for each of the 9 transcript sets. The results of the analysis are shown in Table 1.

	CUHTK	DERA	MDS	SHEF	Mean Cor	Mean Rank
WER	-0.901	-0.905	-0.785	-0.797	-0.847	6.00
SWER	-0.927	-0.912	-0.812	-0.827	-0.869	3.25
NE-WER	-0.937	-0.900	-0.897	-0.890	-0.906	3.00
NE-SWER	-0.936	-0.886	-0.900	-0.898	-0.905	3.00
SWF-WER	-0.894	-0.911	-0.777	-0.791	-0.843	7.00
SWF-SWER	-0.911	-0.915	-0.794	-0.811	-0.858	4.00
SSWF-WER	-0.897	-0.913	-0.776	-0.793	-0.845	6.25
SSWF-SWER	-0.914	-0.916	-0.794	-0.812	-0.859	3.25

Table 1 . Correlation Between Recognition Metrics and Retrieval Performance

The table shows that, on average for all 4 retrieval systems, the named entity test set word error rate (ne-wer) and named entity mean story word error rate (ne-wer) provide the best correlation with retrieval performance with mean system R-squared values of .906 and .905. This is visually depicted in Figure 5 which shows retrieval performance versus recognition performance in terms of named entity mean story word error rate (ne-swer).

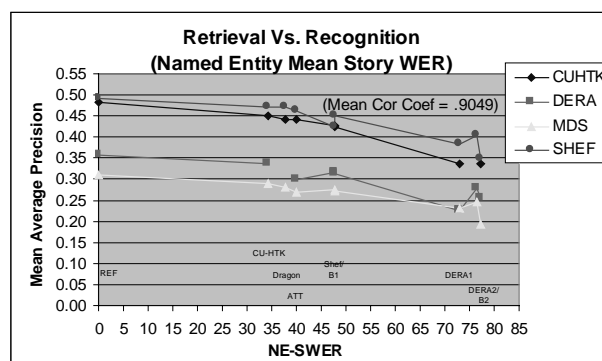


Figure 5. Cross-Recognizer Results: Mean Average Precision vs. Named Entity Mean Story Word Error Rate

5. CONCLUSIONS

In 1998, we found that we could successfully implement an ad-hoc retrieval task using a larger corpus of broadcast news. The best performance for retrieval using a speech recognizer (.5120 MAP) approached the best performance for retrieval using perfect human-generated reference transcripts (.5668).

We also found that there is a near-linear relationship between recognition word error rate and retrieval performance. We also investigated alternative metrics for recognition performance that might be more predictive of retrieval performance. We found that named-entity word error rate was more highly correlated with retrieval performance than word error rate alone.

However, we are hesitant to declare retrieval using speech recognition-generated transcripts a solved problem. The 1998 SDR collection of 2,866 stories is still quite small for retrieval evaluation. The next challenge is to determine whether retrieval performance scales for larger realistic collections of broadcast news and to remove artificially constrained components of the evaluation such as known story boundaries.

6. FUTURE

For 1999, we plan to use the entire TDT-2 corpus for the SDR evaluation. The TDT-2 corpus, which was collected by the Linguistic Data Consortium for the DARPA Topic Detection and Tracking Tasks, contains 632 hours/24,503 stories of broadcast news evenly sampled over a 6-month time period from January through June, 1998. [11]

The NIST assessors will create 50 ad-hoc-style topics for the 1999 SDR track using the existing transcripts. These transcripts will also be used in the reference condition for the evaluation.

There was increased interest at TREC-7 in supporting an evaluation condition in which story boundaries are unknown which would more naturally model a real implementation of SDR. To support this condition, systems will be permitted to make an optional run on the baseline speech recognizer transcripts and their own recognizer output sans story boundaries. The systems will output a time stamp for the top 1000 retrieved stories rather than a story ID. The time stamps will be mapped to reference story IDs prior to standard TREC_EVAL scoring. Duplicate stories will be removed, so systems which over-generate time stamps will be penalized. The goal of this task will be for systems to either find the mid-point or a single topical "hotspot" within relevant stories.

NOTICE

Views expressed in this paper are those of the authors and are not to be construed or represented as endorsements of any systems, or as official findings on the part of NIST or the U.S. Government.

REFERENCES

- [1] Voorhees, E., Garofolo, J., Stanford, V., and Sparck Jones, K., *TREC-6 1997 Spoken Document Retrieval Track Overview and Results*, Proc. TREC-6, 1997 and 1998 DARPA Speech Recognition Workshop, February 1998.
- [2] 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.
- [3] Garofolo, J.S., Voorhees, E.M., Auzanne, C.G. P., Stanford, V.M., Lund, B.A., *1998 TREC-7 Spoken Document Retrieval Track Overview and Results*, Proc. TREC-7, 1998 and 1999 DARPA Broadcast News Workshop, March 1999.
- [4] Johnson, S.E., Jourlin, P., Moore, G.L., Sparck Jones, K., Woodland, P.C., *Spoken Document Retrieval for TREC-7*, Proc. TREC-7, November 1998.
- [5] Allan, J., Callan, J., Sanderson, Xu, J., *INQUERY and TREC-7*, Proc. TREC-7, November 1998.
- [6] Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F., *AT&T at TREC7*, Proc. TREC-7, November 1998.
- [7] 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.
- [8] Miller, D., Schwartz, R., Weischedel, R., Stone, R., *Named Entity Extraction from Broadcast News*, Proc. 1999 DARPA Broadcast News Workshop, March, 1999.
- [9] Douthout, A., *Hub-4 1998 IE-NE Scoring Software with Recognition and Extraction Evaluation Pipeline*, SAIC, ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98_scoring.tar.Z
- [10] Porter, M.F., *An algorithm for suffix stripping*, *Program 14 (3)*, July 1980, pp. 130-137.
- [11] Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., *TDT-2 Text and Speech Corpus*, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

Appendix A: 1998 TREC-7 Spoken Document Retrieval Track Summary Results

Retrieval Results - Mean Average Precision

Site	R1	B1	B2	S1	S2	CR-ATT	CR-CUHTK	CR-DERA1	CR-DERA2	CR-Dragon	CR-Shef
ATT	0.4992	0.4700	0.4065	0.5065	0.5120	0.5065					
CMU1	0.3577	0.3345	0.2590	0.3224	0.2926						
CMU2	0.3936	0.3472	0.2693								
CUHTK	0.4817	0.4272	0.3352	0.4509		0.4419	0.4509	0.3352		0.4428	0.4251
DERA	0.3579	0.3164	0.2551	0.2242	0.2768		0.3375	0.2242	0.2768	0.2990	0.3134
RMIT-MDS	0.3107	0.2753	0.1937	0.0223	0.1063	0.2812	0.2906	0.2309	0.2443	0.2704	0.2730
NSA	0.3907	0.3640	0.2868								
SHEF	0.4916	0.4243	0.3471	0.4495		0.4717	0.4713	0.3836	0.4047	0.4613	0.4495
TNO	0.3970	0.3533	0.2833	0.0436							
UMass	0.5668	0.5063	0.4191	0.5075	0.5000						
UMD	0.4386	0.2557	0.1967								

Speech Recognition Results - Various Metrics (%error)

ASR Metric	R1	B1	B2			CR-ATT	CR-CUHTK	CR-DERA1	CR-DERA2	CR-Dragon	CR-Shef
WER	0.0	33.8	46.6			31.0	24.6	66.0	61.3	29.5	35.6
SWER	0.0	31.9	46.1			27.4	22.2	61.6	58.9	27.3	31.3
NE-WER	0.0	46.8	77.1			40.8	34.2	75.9	75.9	40.6	49.0
NE-SWER	0.0	47.7	77.3			37.8	34.2	73.1	76.5	40.1	47.7
SWF-WER	0.0	35.1	47.2			31.8	25.7	70.0	61.3	30.2	37.0
SWF-SWER	0.0	33.1	45.7			28.4	23.6	65.3	59.0	28.6	32.9
SSWF-WER	0.0	33.2	44.9			29.9	24.0	67.4	58.9	28.3	34.8
SSWF-SWER	0.0	31.3	43.5			26.7	22.1	62.6	56.6	26.8	30.8