# Automatic Language Model Adaptation for Spoken Document Retrieval

**Cédric Auzanne, John S. Garofolo, Jonathan G. Fiscus, William M. Fisher**

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899-8940, USA
{cedric.auzanne, john.garofolo, jonathan.fiscus, william.fisher}@nist.gov

**Abstract**

This paper describes experiments implemented at NIST in adapting language models over time to improve recognition of broadcast news recorded over many months. These experiments were designed specifically to improve the utility of automatically generated transcripts for retrieval applications. To evaluate the potential of the approach, a time-adaptive automatic speech recognition run was implemented to support the 1999 TREC Spoken Document Retrieval (SDR) Track – more than 500 hours of broadcast news sampled across 5 months. The accuracy of retrieval for several systems using the time-adaptive system transcripts was evaluated against transcripts produced by virtually the same recognition system with a fixed language model.

This paper details the process we employed to identify and implement the time-adaptive language model and discusses the results of the experiment in terms of its effect on word error rate, out of vocabulary rate and retrieval accuracy (Mean Average Precision).

## 1. Introduction

Spoken Document Retrieval (SDR) technology enables a user to search an audio collection through typed queries. SDR involves the combination of two human language technologies: automatic speech recognition (ASR) and information retrieval (IR). ASR technology is used to create a searchable time-stamped textual representation of the audio, which can then be indexed and searched using conventional text retrieval algorithms. The time stamps provide pointers into the source audio for playback. The NIST eighth Text REtrieval Conference (TREC-8) Spoken Document Retrieval Track provides an environment for evaluation of such technologies (Garofolo et al, 2000). The SDR Track uses audio recordings of radio and television news broadcasts as its spoken document collection.

Although ASR algorithms and processor speeds have greatly increased in recent years, the ASR component of an SDR system is still likely to require a great deal more processing time than the IR indexing component. These processing times are likely to be significant for realistically large collections. For example, with state-of-the-art 10xRT ASR systems (e.g. Davenport et al,1999; Odell et al, 1999; Wegmann et al, 1999; Sankar et al, 1999), the processing time for a 500-hour collection would require 5000 processor hours. This time is significant, even when using a multiple-processor approach.

Therefore, in a real working implementation of an SDR system, the speech recognition portion of the task would have to be performed *online*, that is, as the speech data is collected and made available, rather than *retrospectively*, after all of the audio has been recorded. This means that the recognizer cannot use future data to optimize its models nor can the audio be re-transcribed later on with better models. It makes sense to use the best possible recognition algorithm for each recognition run. In contrast, the derived textual form of the entire collection can be completely re-indexed on a daily basis.

Current HUB-4 style (Pallett et al, 1998) speech recognizers use static, pre-trained recognition algorithm models. If such a recognizer is used on a collection that spans several months of broadcast news, the system will be unable to recognize new words as they appear in the news -- words that are

likely to be important for the retrieval engine[1]. The recognition models and the actual spoken language will gradually diverge over time, eventually yielding significant increases in word error rate.

We hypothesized that we could continuously update the language model of a recognizer by using a parallel text corpus, which contained approximately the same language characteristics as the spoken corpus to be recognized. Given that we were working in the broadcast news domain for SDR, we believed that a contemporaneous newswire text corpus could be used for the adaptation. The broadcast news domain provides a good testbed for our hypothesis since as news events change, new words appear gradually over time. Further, these new words are also likely to appear almost immediately in collateral newswire texts. Thus, the newswire data could be used to perform a periodic update of the recognizer dictionary, which contains the list of words that the speech recognizer understands and their pronunciations. This updating would lower the discrepancy between the recognizer model and the spoken language. In an online recognition scenario, this means using today's (and all previous days') news to hypothesize tomorrow's lexicon.

Since the TREC-8 SDR corpus contained broadcast news sampled over a 5 month period (Garofolo, et al., 1999), it provided us with the perfect testbed for experimentation with such adaptive language modeling techniques. The work we report on here is similar to other recent work on cache language modeling. (Clarkson et al, 1997, Kuhn et al, 1990, 1992).

## 2. The TREC SDR track

For the past three years, NIST has organized the SDR track and provided the evaluation infrastructure for the task by providing test specifications, scoring software, speech and text corpora, and test topics. For the past two years, NIST has also created a "baseline" set of recognizer transcripts using a contributed research recognizer. These baseline transcripts have provided a valuable control condition as well as permitted sites without access to their own speech recognizers to participate in the evaluations[2] (Garofolo et al, 1999, 2000).

### 2.1. The test corpora

The 1999 TREC-8 SDR collection is based on the broadcast news audio portion of the TDT-2 News Corpus (Cieri et al 1999) which was originally collected by the Linguistic Data Consortium (LDC) to support the DARPA Topic Detection and Tracking Evaluations (Fiscus et al., 1999). The corpus contains recordings, transcriptions, and associated data for several radio and television news sources broadcast daily between January 04 and June 30 1998. In total, TDT-2 contains 1,064 broadcasts (shows) and 657.5 hours of digitally sampled audio.

The 1999 SDR Track used the February - June subset of the TDT-2 corpus (The month of January was excluded so as not to conflict with Hub-4 recognizers which had been trained on overlapping material from January 1998). The resulting SDR test collection contained 902 broadcasts and 557.5 hours of audio.

Unlike previous SDR corpora, the 1999 collection did not contain detailed human-generated transcriptions especially created for ASR training and testing. Rather, it contained "found" closed-caption transcriptions for the television sources and paid for transcriptions for the radio sources. These found transcripts were used as the reference form for retrieval purposes. However, in order for

[1] Important words for retrieval are those likely to appear in a query, but in a real production system, virtually any word in the documents might appear in a query, as one might expect up-to-date queries.

[2] The test specifications and document regarding the 1999 SDR Track are available at http://www.nist.gov/speech/sdr99/sdr99.htm. Information regarding the upcoming 2000 SDR Track will be made available at http://www.nist.gov/speech/sdr2000/sdr2000.htm.

us to more accurately benchmark the participants' ASR transcripts, the LDC transcribed a 10-hour randomly selected subset of the corpus to provide Hub-4[3] training quality transcripts.

The TDT-2 corpus also provided newswire texts, which runs parallel to the recorded broadcast news. This text data was made available to SDR participants for adaptive language modeling purposes. Table 1 provides a breakdown of the corpus by source.

| Source | Number of sentences | Number of tokens | Number of types |
|---|---|---|---|
| **Associated Press** | 1 493 708 | 31 482 921 | 199 263 |
| **New York Times** | 818 439 | 17 289 108 | 135 470 |
| **TOTAL** | 2 312 147 | 48 772 029 | 258 243 |

Table 1 - SDR 99 newswire collection distribution

### 2.2. The baseline (B1) recognizer transcript set

As in TREC-7, NIST wanted to provide a baseline recognizer transcript set with which to compare retrieval performance across sites and which permitted retrieval sites without speech recognition technology to participate in the evaluation. Realizing that the CMU SPHINX recognizer was far too slow to recognize the TREC-8 collection in the time allotted, NIST set out to find a faster baseline recognizer. To our good fortune, the 1998 Hub-4 evaluation included an evaluation of fast speech recognizers – systems that operated in under 10 times real time on a single processor. The new evaluation had caused several ASR sites to improve the speed and efficiency of their recognizers. GTE/BBN offered NIST a Linux instantiation of their fast BYBLOS Rough 'N' Ready (BBN-RNR) recognizer (which now operated at only 4xRT) to use as a baseline in SDR and TDT (Davenport et al., 1999; Kubala et al, 2000). Before experimenting with adaptation, NIST decided to create a control recognizer (B1) using a traditional Hub-4-style fixed language model.

#### 2.2.1. Recognizer description

The baseline recognizer transcripts were created at NIST using the BBN-RNR system trained with the following data:

- Acoustic model training: 1995 (Marketplace) and 1996/97 (Broadcast News) training sets released by the Linguistic Data Consortium for use in Hub-4 speech recognition evaluations
- Language model training: a set of LDC-published text sources totaling 480 M words
- Dictionary: Original BBN dictionary with 50,929 carefully chosen words and their pronunciations

In performing these runs, we benchmarked BBN-RNR at less than 4 times real time using state-of-the-art hardware of the time (Pentium II 450 Mhz with 512 MB of RAM running Linux 2.0.36). The results on previous HUB-4 evaluation test sets, given in Table 2, are similar to those obtained at BBN (Kubala, et al., 2000).

| Test | % WER | x RT |
|---|---|---|
| **Eval 96** | 30.1 | 3.5 |
| **Eval 97** | 22.7 | 3.5 |
| **Eval 98_1** | 22.8 | 3.2 |
| **Eval 98_2** | 19.9 | 3.6 |

Table 2 – BBN-RNR performance on past HUB-4 test sets

#### 2.2.2. Performance

It took approximately 3 weeks to run the recognizer on the entire SDR corpus using our 16-processor PC-based cluster. The open source PBS[4] scheduler was used to distribute the processing of the data

[3] Hub-4 is a DARPA/NIST-sponsored evaluation of ASR technology (Pallett et al, 1998).
[4] Portable Batch System is available from http://pbs.mrj.com/

over the processors. It also provided logging and error checking. The results on the 10-hour scoring subset for the TREC-8 SDR track are given in Table 3.

| | # Snt | # Words | % Corr | % Sub | %Del | % Ins | % Err | % Sent. Err |
|---|---|---|---|---|---|---|---|---|
| **NIST-B1** | 1 945 | 88 156 | 75.8 | 17.8 | 6.4 | 3.3 | 27.5 | 91.8 |

Table 3 - NIST-B1 scoring results

In the following, all the out-of-vocabulary (OOV) rates are computed on word frequency list files extracted from the human-generated closed captioning. The closed-caption is not as accurate a transcript as the usual HUB-4 transcripts provided by the LDC. Hence, errors and ASR/closed-caption spelling mismatch must be taken in account when reading the OOV rates. Nevertheless, this permitted us to compute OOV rates on a much larger corpus than the 10-hour subset.

To illustrate the discrepancy between closed-caption quality and Hub-4 training quality, consider the OOV rates computed using the original BBN dictionary for the same 10-hour subset:

- on Hub-4 training quality transcripts: 2.04 %
- on closed-captioning quality transcripts: 2.71 %

These differences are due in part to lack of quality assurance in closed captioning and also because that closed-caption transcripts are not normalized for ASR. For instance “$5” should be written “five dollars”. The OOV numbers in this article should be seen as pessimistic.

Using the recognizer’s dictionary, we found that the out-of-vocabulary (OOV) rate for the recognizer with regard to the entire SDR corpus was 2.54 %.

## 3. Implementing a Rolling Language Model

We found in TREC-7 that retrieval accuracy was more highly correlated with content word recognition accuracy than on simple recognition word error rate (Garofolo et al, 1998). For our broadcast news domain, we hypothesized that a reduction in OOV rate would increase retrieval accuracy since new words in the news are likely to be content-bearing words which are supposed to be important for retrieval. This year, we set out to create a second (B2) experimental baseline recognizer using an adaptive language model which would give a lower OOV rate than a comparable recognizer using a fixed language model does.

We identified several possible parameters we could vary in implementing continuous language model adaptation:

- Lexicon size balancing (word addition/deletion)
- Pronunciation model generation
- Language model/dictionary update period
- Language model retraining method

### 3.1. Lexicon Balancing

To keep the first experiment tractable and produce a second recognizer run in time for use in the SDR track, we employed a technique that augmented the words in the original BBN language model. The original BBN dictionary contained 50,929 words. This left a margin of 14,000 words which could be added before reaching the 64K-word limit recommended by BBN.

The process of word selection attempts to predict which words will be used in the language. One option is to add all the words encountered. However, such approaches inevitably fail because the lexicons grow so large that the language model perplexity grows out of control. We therefore need to judiciously choose only the words that have a fairly high probability of occurring.

One approach to limiting the number of word additions is to look at the number of occurrences of a new word in the training data (newswire texts) and set a threshold for additions and deletions. In this

approach, a word is added if is encountered a minimum of *count* times per day. Likewise, a new word is removed from the lexicon if it has not been encountered for *lookback* days. To select these parameters, we created a tool that simulates such a language model update procedure and computes the OOV rate based on different parameter values. We ran a set of experiments using the first two months of the newswire corpus (January and February 1998) as LM training data and computed the OOV rate on the same two months of closed-caption reference texts. For the lookback period, we used 7 to 28 days in increments of 7. For the count, we used 1 to 10 occurrences in increments of 1. We found that a lookback of 28 days and a count of 4 minimized the OOV rate for our collection.

### 3.2. Pronunciation Model Generation

To generate the necessary pronunciations for new words to be added to the dictionary, we used a statistical text-to-phone (TTP) engine created by William M. Fisher (Fisher, 1999) which has a segmental phonemic accuracy of 94.5% when trained and tested on an English pronunciation dictionary. The pronunciations of the original words in the BBN dictionary were left as-is.

### 3.3. Language Model/Dictionary Update Period

In choosing an adaptation frequency, we had to balance the utility of using the most up-to-date language model against the computational cost of producing one. The process of completely updating the language model took approximately 6 hours on a Linux workstation. We chose initially a 7-day update period since it would capture the weekly news program cycle and maintain a reasonable computational cost. We also wanted to store each of the adapted language models for further study[5] We, therefore, ended up generating 22 weekly language models spanning February 1st to June 30th.

### 3.4. Language Model Retraining

For expediency, we used the same newswire data to retrain the language models as we used to adapt the dictionary and lexicon. This data was added to a history count maintained by the BBN language model tools. It is well known that language model training generally requires large quantities of text in order to properly sample and model the language. While our parallel news corpus was probably adequate for finding new words, it may have been inadequate to fully train the language model for those new words. Therefore, our word-to-word probabilities are questionable. However, we made the best possible use of the data we had.

### 3.5. System Configuration

The system we used to implement language model adaptation is shown in Figure 1. It begins with the newswire text corpus from which the set of new words to be added to the lexicon is derived using the Selection Tool. The TTP tool is then used to create pronunciation models for the new words to be added to the dictionary. Entries from the original dictionary are kept as is. Hence, the updated

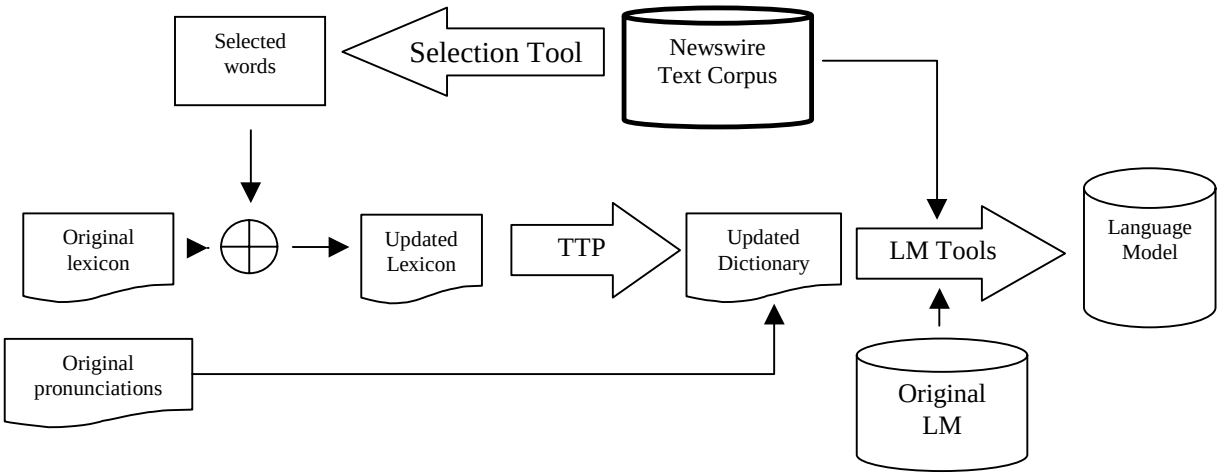


Figure 1 – System diagram

[5] The storage requirements (~600Mb per model) for daily models would have been prohibitively large.

dictionary contains all the original words and pronunciations plus the new words with the automatically generated pronunciations. The newswire texts are then used with the revised dictionary to create an updated language model.

Table 4 summarizes the resulting OOV rate relative to the closed-captioning data for three different conditions: fixed (B1) language model, adaptive (B2) language, and a minimum OOV case in which all new words are added and never discarded. Two OOV rates are shown: one for a two-month period, and the second for a one-month period. The one-month result is more informative since it has a full 28-day lookback for all days.

| | Fixed LM (nist-B1) | Rolling LM (nist-b2) | Best case |
|---|---|---|---|
| **%OOV on jan-feb 1998** | 2.68 | 2.14 | 1.82 |
| **%OOV on feb 1998** | 2.71 | 1.97 | 1.55 |

Table 4 - Average OOV rate over partial corpus

Figure 2 shows a plot of the daily OOV rate for each of the 3 conditions. The graph shows no divergence for the first seven days since none of the models are updated yet. For the next 21 days, the models begin to diverge as the lookback window grows to the maximum possible 28 days. The

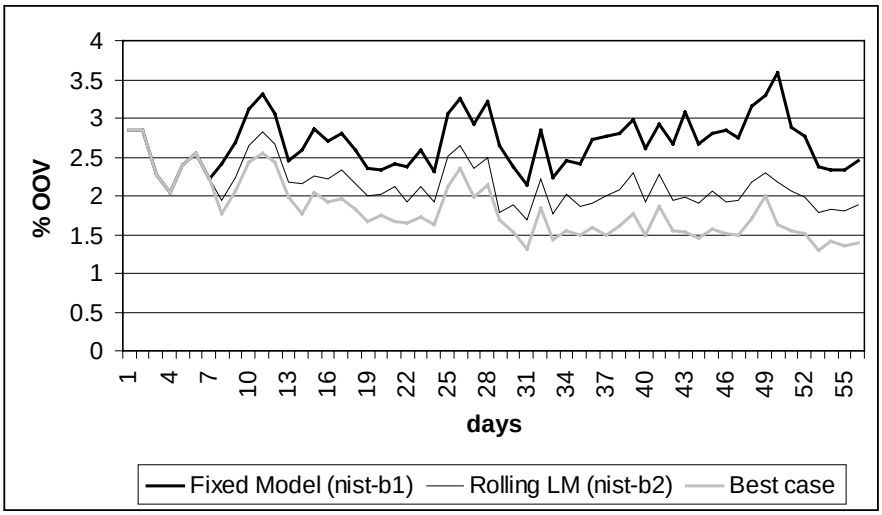


Figure 2 - OOV rate comparison for 3 models for january-february

difference in OOV for the next 28 days appears to be stable. This is somewhat counter-intuitive, as one would expect continued divergence between the adaptive and fixed models. The language is actually very stable for this short sample period and the OOV rate, therefore, appears to be constant for the fixed model. This might not be the case for a much longer sample in which the core words in the language begin to change.

## 4. Adaptive Recognition Experiment Results

With the parameters set to a lookback period of 28 days, a minimum count of 4, and update period of 7 days, we created the 22 language models for the complete SDR corpus. For the B2 recognizer, we found that the OOV rate relative to the closed-captioning data for the entire test corpus was 1.97 %. We then ran the recognizer over the corpus and scored the results using the 10-hour subset (Table 5).

| | # Snt | # Words | % Corr | % Sub | %Del | % Ins | % Err | % Sent. Err |
|---|---|---|---|---|---|---|---|---|
| **NIST-B1** | 1 945 | 88 156 | 75.8 | 17.8 | 6.4 | 3.3 | 27.5 | 91.8 |
| **NIST-B2** | 1 945 | 88 156 | 76.5 | 17.2 | 6.2 | 3.2 | 26.7 | 91.5 |

Table 5 - NIST-B2 ASR scoring results

The results show a 0.8% absolute decrease in word error rate between the adaptive B2 system and the fixed B1 system, representing a 2.9% relative decrease in word error rate. Although seemingly small, the NIST statistical significance software indicates that this difference is significant. Not surprisingly, the insertion and deletion rates are similar for the two recognizers. However, the substitution rate for the B2 recognizer is 0.7% lower than that for the B1 recognizer. This indicates that previously-unrecognizable words are now being correctly recognized because of the adapted lexicon and language model.

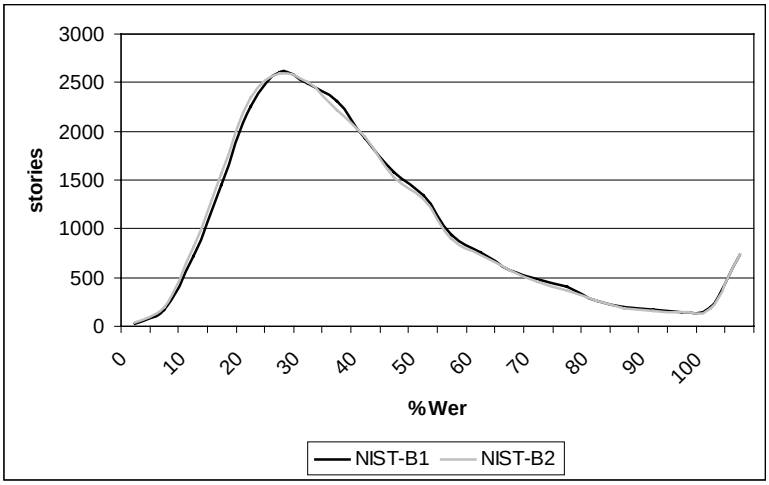


Figure 3 – B1 / B2 WER histogram comparison

The performance of the two recognizers are visually depicted in Figure 3, which shows a histogram of number of stories against word error rate. The graph for the B2 recognizer is shifted to the left and flatter than that of the B1 recognizer – indicating that a majority of the stories benefited from the gain. Table 6 summarizes the OOV rate for the two baseline recognizer runs over the entire TDT corpus and over the SDR test set (TDT subset). The results show that the overall OOV rate for the entire 5 months SDR corpus is a bit less than that of the subset we based our estimations, on indicating that the first two months of the corpus contain a greater percentage of OOV words. This could be explained by the lack of Voice of America broadcasts in the first two months. Presumably, since the VOA broadcasts are geared toward non-native speakers, the language in it is simpler than the other sources that were broadcast for primarily American audiences. The more interesting point, however, is that the adaptive language model provided a 22.44% relative reduction in OOV for the SDR corpus.

| | **Fixed LM (nist-B1)** | **Rolling LM (nist-b2)** | **Relative reduction (B2 – B1)** |
|---|---|---|---|
| **%OOV on jan-june 1998** | 2.56 | 2.02 | 21.09 |
| **%OOV on feb-june 1998** | 2.54 | 1.97 | 22.44 |

Table 6 - Average OOV rate for the Baseline runs for the complete SDR corpus

Figure 4 shows a daily graph of the relative reduction in OOV between the B2 and B1 recognizers. The graph shows a 7-day spike in OOV, which is probably related to a weekly broadcast news cycle effect. However, there also seems to be a larger periodicity that could be explained by seasonal variations in the news.

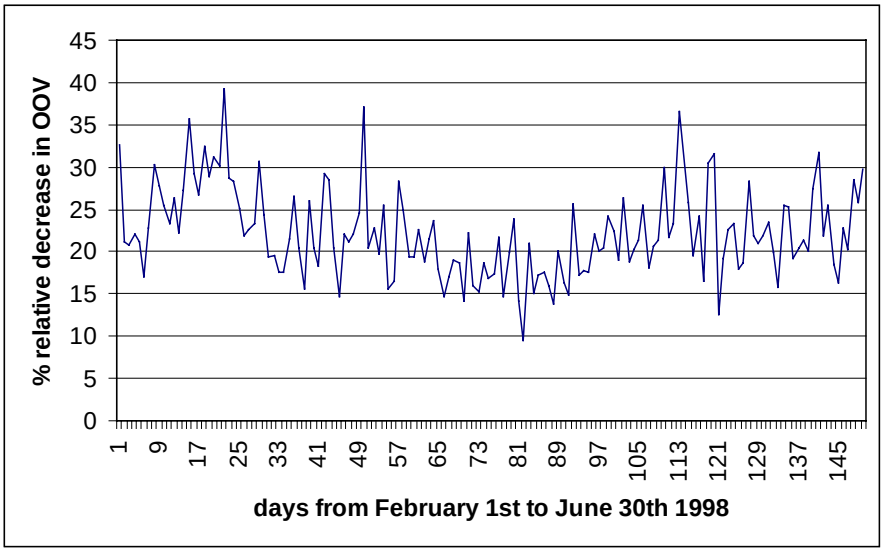


Figure 4 - Relative gain in OOV for B1 /B2

### 4.1. Effect of Adaptive Recognition on Retrieval

Our hypothesis was that an adaptive recognizer with fewer OOV words would also produce better retrieval results. Three of the TREC-8 SDR participants ran their retrieval engine on both the B1 and B2 recognition transcripts: Cambridge University (Johnson et al, 2000), LIMSI (Gauvain et al, 2000) and Sheffield University (Abberley et al, 2000). Table 7 summarizes the results for the SDR retrieval test for each of the test sites and baseline recognizers (Garofolo et al, 2000). The retrieval metric used is mean average precision (MAP), the standard metric used in TREC ad-hoc tests (Voorhees et al 2000).

| Site | Baseline 1 | Baseline 2 | % relative gain |
|---|---|---|---|
| **CU-HTK** | 0.5281[6] | 0.5302 | 0.39 |
| **LIMSI** | 0.4828 | 0.4839 | 0.22 |
| **Sheffield** | 0.5298 | 0.5335 | 0.69 |

Table 7 - IR results for the two baseline runs

The results indicate a small, but consistent gain in retrieval accuracy for the B2 recognizer. The small gain is indicative of previous results we have found which demonstrate that retrieval performance is minimally effected by recognition performance. This is mainly due to the redundancy of content words in the collection (Garofolo, et al., 2000).

## 5. Alternative Adaptation Experiments : Frequency-based Algorithm

Some words in the news may appear a limited number of times per day, but appear with a consistent frequency. The minimum count algorithm used in the B2 recognizer would not pick up these words. We hypothesize that the algorithm could be improved if we add a frequency component. To implement this, we would add a word to the language model if it appeared *count* days over the past *window* days.

We performed a similar OOV experiment as above varying the frequency while using the optimal lookback period of 28 days and count of 4 we determined earlier. We varied the frequency window

[6] The official result as published in the TREC paper is 0.4963 for the CU-HTK retrieval engine on the NIST baseline 1 recognizer transcript. However, this result was due to a bug in the procedure and the actual number is 0.5281 (Johnson et al, 1999).

from 7 to 28 days with 7-day increments and the frequency count from 1 to 6 with increments of 1. We found that a frequency window of 28 days and a frequency count of 4 gave us the lowest OOV rates.

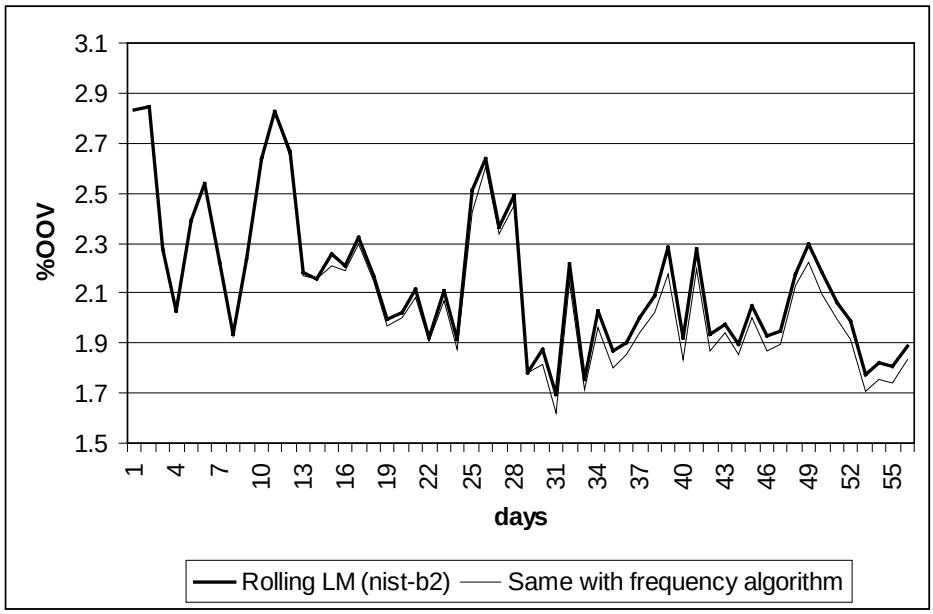


Figure 5 – OOV rate comparison for B2 algorithm and new frequency based algorithm

Figure 5 shows the daily OOV rate for the new and original B2 algorithms. Although the reduction is insignificant for the first month while the frequency window covers less than the full 28 days, there is a small, but measurable reduction (3% relative) for the second month.

## 6. Conclusion

The 1999 TREC-8 SDR track, with its large audio collection and collateral newswire corpus provided a nearly ideal testbed for the exploration of automatic language model adaptation. We found that we could significantly improve the OOV rate and word error rate for the BBN BYBLOS Rough 'N' Ready recognizer on the SDR test corpus by implementing a time- and count-based adaptation algorithm using parallel newswire texts to augment the recognizer's language model. The improved recognition translated into a small improvement in retrieval performance for 3 retrieval engines. We have begun to explore additional approaches to further improve the recognizer's OOV performance including making use of word frequency information in selecting words to be added to the recognizer's language model.

## 7. Acknowledgements

NIST's work in the TREC-8 SDR track has been sponsored, in part, by the Defense Advanced Research Projects Agency.

We would like to thank Francis Kubala and Daben Liu from BBN/GTE for the generous donation of their BYBLOS Rough 'N' Ready recognizer and language modeling tools for our use in developing the SDR baseline ASR transcripts and for their support in helping us to install, configure, and use the software.

## 8. References


Abberley D., Renals S., Robinson T., Ellis D. (2000). The THISL SDR System At TREC-8. In *Proceedings of the Eighth Text REtrieval Conference*. 2000.

Cieri C., Graff D., Liberman M., Martey N., Strassel S. (1999). The TDT-2 Text and Speech Corpus. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

Clarkson P. R., Robinson A. J. (1997). Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of ICASSP*, 1997.

Davenport J., Nguyen L., Matsoukas S., Schwartz R., Makhoul J. (1998). The 1998 BBN BYBLOS 10x Real Time System. In *Proceedings of the 1999 DARPA Broadcast News Workshop.* 1999.

Fiscus J., Doddington G., Garofolo J. S., Martin A. (1999). NIST's 1998 Topic Detection and Tracking Evaluation (TDT-2). In *Proceedings of the DARPA Broadcast News Workshop,* 1999.

Fisher W. M. (1999). A statistical Text-To-Phone function using Ngrams and rules. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 1999.

Garofolo J. S., Auzanne C., Voorhees E. M. (2000). The TREC Spoken Document Retrieval Track : A Success Story. To appear in *Proceedings of RIAO-2000.*

Garofolo J. S., Voorhees E. M., Auzanne C. G. P., Stanford V. M., Lund B. A. (1998). 1998 TREC-7 Spoken Document Retrieval Track, Overview and Results. In *Proceedings of the 1999 DARPA Broadcast News Workshop*. 1999.

Gauvain J. L., de Kercadio Y., Lamel L., Adda G. (2000). The LIMSI SDR System for TREC-8. In *Proceedings of the Eighth Text REtrieval Conference.* 2000.

Johnson S. E., Jourlin P., Sparck Jones K., Woodland P. C. (2000). Spoken Document Retrieval for TREC-8 at Cambridge University. In *Proceedings of the Eighth Text Retrieval Conference.* 2000.

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

Kuhn R., De Mori R. (1990). A cached based Natural Language Model for Speech Reproduction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12(6):570-583, 1990.

Kuhn R., De Mori R. (1990). Corrections to 'A cached based Natural Language Model for Speech Reproduction'. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 14:691-692, 1992.

Odell J. J., Woodland P. C., Hain T. (1999). The CUHTK-Entropic 10xRT Broadcast News Transcription System. In *Proceedings of the 1999 DARPA Broadcast News Workshop,* 1999.

Pallet D. S., Fiscus J. G., Garofolo J. S., Martin A., Przybocki M. A. (1998). 1998 Broadcast News Benchmark Test Results. In *Proceedings of the 1999 DARPA Broadcast News Workshop,* 1999.

Sankar A, Gadde R. R., Weng F. (2000). SRI's 1998 Broadcast News System - Toward Faster, Better, Smaller Speech Recognition. In *Proceedings of the Eighth Text REtrieval Conference.* 2000.

Voorhees E., Harman D. (2000). Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference.* 2000.

Wegmann S., Zhan P., Carp I., Newman M., Yameon J. P., Gillick L. (1999). Dragon Systems' 1998 Broadcast News Transcription System. In *Proceedings of the 1999 DARPA Broadcast News Workshop*. 1999.