Document Expansion for Speech Retrieval

Amit Singhal Fernando Pereira
AT&T Labs — Research
http://www.research.att.com/info/{singhal, pereira}

Abstract

Advances in automatic speech recognition allow us to search large speech collections using traditional information retrieval methods. The problem of "aboutness" for documents — is a document about a certain concept — has been at the core of document indexing for the entire history of IR. This problem is more difficult for speech indexing since automatic speech transcriptions often contain mistakes. In this study we show that *document expansion* can be successfully used to alleviate the effect of transcription mistakes on speech retrieval. The loss of retrieval effectiveness due to automatic transcription errors can be reduced by document expansion from 15–27% relative to retrieval from human transcriptions to only about 7–13%, even for automatic transcriptions with word error rates as high as 65%. For good automatic transcriptions (25% word error rate), retrieval effectiveness with document expansion is indistinguishable from retrieval from human transcriptions. This makes speech retrieval from automatic transcriptions, even poor ones, competitive with retrieval from perfect transcriptions.

1 Introduction

Increasing amounts of spoken communication are stored in digital form for archival purposes (for instance, broadcasts) or as a byproduct of modern communications technology (voice-mail for instance). Furthermore, multimedia documents and databases are becoming increasingly popular, for example on the Web. It would be therefore desirable to have tools searching spoken information that complement the existing methods for searching textual information.

With advances in automatic speech recognition (ASR) technology, it is now possible to automatically transcribe speech with reasonable accuracy [17]. Once the contents of a speech database or the audio portions of a multimedia database are transcribed using a speech recognition system, traditional information retrieval techniques can be used to search the database. However, inaccuracies in the automatic transcriptions pose several new problems in use of IR technology for speech retrieval.

Even though IR techniques have been successfully used in retrieving corrupted text generated by optical character recognition (OCR) systems [26], the kinds of errors in automatic speech transcriptions are very different from those in OCR transcriptions. Since OCR systems usually operate with single characters, errors in character recognition usually produce illegal words which do not affect the retrieval process substantially. Whereas current high-performance large-vocabulary speech recognizers rely on word-pronunciation dictionaries, and their output consists only of legitimate words, drawn from the dictionary. Recognition errors are then deletions, insertions or substitutions of legitimate words, and are therefore not easily discarded.

In this study we explore document expansion for improved speech retrieval. The motivation for this approach is discussed in detail in Section 2. We explore various parameters for document expansion and show that document expansion yields significant improvements in retrieval effectiveness from speech. When speech recognition is of reasonable quality, retrieval effectiveness from automatic transcriptions is quite comparable with retrieval effectiveness from human transcriptions. However, for poor automatic transcriptions, retrieval effectiveness is much worse (about 15–27% worse) than effectiveness for human transcriptions. Due to various factors, including background non-speech sounds (noise, music), poor recording conditions, and disfluent or non-native speech, it is often not possible to get good automatic transcriptions even with the best ASR systems. Therefore, techniques that improve effectiveness of retrieval from poor transcriptions are urgently needed. We find that document expansion is very effective exactly in that situation. The losses in retrieval effectiveness are reduced considerably from 15–27% to about only 7–13%.

The rest of this study is organized as follows. Section 2 puts the problems in indexing of automatic speech transcriptions in perspective. Section 3 reviews previous studies in speech retrieval. Section 4 describes the experimental environment. Section 5 details our experiments and discusses their results. Section 6 explores the effect of document expansion from a corpus that is not well related to the speech corpus. Finally, Section 7 concludes the study.

2 "Aboutness" and Vocabulary Mismatch

The main problem in doing word- and phrase-based speech retrieval arises due to poor index term assignments for automatic speech transcriptions. From its early days, the field of IR has wrestled with the question of what index terms should be assigned to a given document [1, 23]. Defining the concepts which a document is about, or "aboutness" in subject indexing, has been visited several times over the history of IR [8]. Experimentation has shown that automatically-derived uncontrolled index terms are competitive with carefully crafted manual index terms [23]. Most modern IR systems use automatically derived words and phrases as index terms for documents. However, any indexing system, including word- and phrase-based automatic indexing, is imperfect and may thus fail to index the relevant documents under the query terms even though the documents are about those terms. This has often been called the "vocabulary mismatch" problem. This problem is obviously made worse by speech recognition errors, since the automatic transcription of a document may not contain all the terms that were actually spoken or may contain terms that were not spoken.

A secondary problem in index term assignment is deciding, for an index term assigned to a given document, the "degree" to which that document is about that term. Modern IR systems use sophisticated term-weighting methods to define the degree of aboutness of documents for different terms [18, 24]. When documents are corrupted, as is the case in speech retrieval, term-weighting schemes assign misleading weights to terms. This might also cause some loss in retrieval effectiveness.

Many devices have been proposed over the years to attack the vocabulary mismatch problem, most notably the use of thesaurii to enhance the set of index terms assigned to documents or to the queries [13]. However, obtaining a reliable thesaurus for any subject area is quite expensive. Attempts have been made to harness word-word associations for automatic thesaurus construction, but these attempts have mostly been disappointing [12]. More recently, however, it has been shown that enhancing queries with terms related to the entire concept of the query (often refered to as *query expansion*), and not just with words related to individual query words, does reduce the problem of vocabulary mismatch considerably and consistently yields large improvements in retrieval effectiveness, especially for short queries [16, 10, 28].

Correspondingly, document expansion can be used to enhance the index term assignment for documents. Many studies have enhanced document representations using bibliographic citations and references [21, 14, 5, 3]. Research on using spreading activation models in IR also aims at crediting documents based on activation of related documents [2, 3]. However, both these technique need some human supervision (in form of human generated citations, or the semantic net used) to be made operational. Document clustering, which doesn't require any human supervision, can also be interpreted as a form of document expansion. When similar documents are clustered and a cluster representative is used in the search process, the cluster representative usually contains terms from all the documents in the cluster, in effect allowing a match between a document and a query (via the cluster representative) even when individual query terms might be missing from the document (but are present in other documents in the cluster). Extensive studies on document clustering have given results that are negative to mixed at best [9, 22, 6, 27, 30]. Work on Latent Semantic Indexing (LSI) also has a similar feel. LSI allows a match between queries and documents that might not share any terms in word-space but do share some concepts in the LSI-space [4].

Even though enhanced document representations (what we are calling document expansion) have given mixed results for perfect text, we hypothesize that it would be beneficial in retrieval from erroneous texts such as automatic speech transcriptions. With erroneous transcriptions, we are unsure whether the document is truly about the terms returned by the recognizer. However, if we can find documents that are topically related to the spoken document in a textual *expansion corpus*, we can reinforce words whose presence in speech is supported by topically related documents, and similarly we can reduce the importance of the words whose presence in the speech document is not supported by related documents. In addition to this reweighing of recognized terms, we can also add to the transcription some new terms from the related documents, representing in-topic terms that could have been spoken but the recognizer failed to recognize. In this paper we evaluate the hypothesis that both reweighing and addition of related terms alleviates the retrieval failures due to ASR errors. These two techniques should be especially effective when the expansion corpus is closely related to the spoken documents, for example in the case of a spoken corpus of news broadcasts and an expansion corpus of newspaper news articles of the same time period.

3 Related Work

Some early work on word-based speech processing that resembles information retrieval methods was done by Rose et al. using word-spotting [20]. More extensive work on word-spotting based speech retrieval was done in the Video Mail Retrieval (VMR) project at Cambridge University [11].

An alternative to word-based approaches is to recognize sub-word units (for instance, phones) and use sequences of these sub-word units as index terms [29]. However, it is unclear if the results from this approach are competitive with word-based approaches now that very-large vocabulary recognition systems are available. It is also possible to use simultaneously as index terms words from the best word transcription and phonetic n-grams from phone lattices [11, 31]. A comparison of the indexing effectiveness of various sub-word units is done in [15].

Although we experiment only with word-based systems in the present study, the techniques explored here are quite general. The methods we propose would enhance the index representation for speech documents independently of the nature of the indexing units. The only precondition is that the speech documents to be searched must be transcribed and indexed, and possibly expanded, in advance of query processing, rather than being transcribed and scanned at query-time.

4 Experimental Setup

We use the TREC-7 SDR track as our experimental testbed [28]. The speech collection for this task is made up of approximately 100 hours of radio/TV broadcast news recordings. These recordings have been manually segmented into 2,866 different stories. Twenty three sentence-length queries are supplied with this collection, along with their corresponding relevance judgments. These queries have anywhere between one to sixty relevant documents in the collection.

We use non-interpolated average precision to evaluate retrieval effectiveness. However, average precision is quite volatile for queries that have very few relevant documents. For example, consider the SDR track query (number 71) which has just one relevant document. If one system ranks the relevant document at rank 1 and another at rank 3, then

Code	Provided By	WER
Human	NIST	0%
CUHTK-S1	Cambridge University	24.8%
Dragon98-S1	Dragon Systems	29.8%
ATT-S1	AT&T Labs	31.0%
NIST-B1	Carnegie Melon (CMU)	34.1%
SHEF-S1	Sheffield University	36.8%
NIST-B2	Carnegie Mellon (CMU)	46.9%
DERASRU-S2	DERA	61.5%
DERASRU-S1	DERA	66.2%

Table 1: Transcriptions used in this study.

for this query, the average precision for the first system is 1.0 whereas it is just 0.33 for the second system. Such large differences for a few such queries can overshadow the overall average precision. To avoid this volatility, we remove four queries that have fewer than five relevant documents in the test collection, and do our evaluation using the remaining *nineteen queries*.

NIST has made available the human transcriptions for this speech, and several track participants have also provided their automatic transcriptions to researchers for experimentation. To study the impact of document expansion on a wide variety of speech recognitions of varying accuracy, we use nine different sets of transcriptions in this work. Table 1 lists the various transcription sets along with their word error rates (WER)[1].

Query Creation

We use the term weighting methods described by Singhal et al. in [25], and use *dtn*-weighted queries (see Table 1 in [25]) in our experiments. These queries incorporate the *idf*-factor which is collection-dependent. As the main aim of this study is to investigate document expansion, we want to hold the queries constant across various representations of documents that we explore here. For this reason, we use the same *idf* for query terms across different sets of transcriptions; this *idf*-factor is the true *idf* of the term derived from the collection of human transcriptions of the speech data.

We also investigate the effects of document expansion for query sets of varying richness. It is well-known that long queries, which are rich with content words, yield better retrieval. Many techniques that are effective for short queries (like automatic query expansion) are not as effective for long queries. Therefore we experiment with one set of short queries, and another set of long queries. However, the queries provided by NIST for this collection are all sentence length, and are relatively short (average 7.1 terms per query). We generate a content-rich long version of these queries via pseudo-feedback on human transcriptions. Pseudo-feedback has been used by many participants at TREC and has been quite successful over the last few years [28]. We add ten new words to each query increasing the average query length to 17.1 terms/query.

5 Experiments and Results

From an IR system's perspective, a recognizer makes three kinds of mistakes for a document. 1) *Deletions*: No occurrence of a term occurring in the speech is recognized. 2) *Weight Difference*: A term in the spoken document is recognized, but with the wrong frequency, or the transcription length is incorrect.(Possibly *idf* for terms is poor but we ignore that in this study.) And 3) *Insertions*: A term not in the speech appears in the automatic transcription.

We use *dnb* weighted documents in our experiments (see Table 1 in [25]). Due to the double-log normalized *tf*-factor used in the document weighting scheme, we believe that the effect of weight difference on retrieval effectiveness is minimal, a fact confirmed in the following experiments.

5.1 Effect of Recognition Mistakes

We first study the effect of various kinds of mistakes that a recognizer makes—deletions, weight difference, insertions—on retrieval effectiveness for speech. We compare the performance of retrieval from various speech transcripts to that of retrieval from human transcriptions. To study the incremental loss due to word deletions, weight difference, and insertions, we take the *dnb*-weighted (see Table 1 in [25]) document vectors for human transcriptions, and do the following in increments:

1. We first remove from the document vectors for human transcriptions, all terms that are not recognized by the recognizer. Comparing retrieval from these truncated vectors to that from the full vectors would measure the effect of deletions on retrieval effectiveness.

2. We then change the weights of the terms in the truncated vectors generated in step 1 to the weights they get in the indexed versions of the automatic transcriptions. Retrieval on these vectors would measure the additional loss in retrieval effectiveness caused by incorrect weights being assigned to terms.

3. Finally, we add all insertions to vectors generated in step 2. This yields the final retrieval effectiveness that we expect to see using automatic transcriptions, and also measures the incremental loss due to insertions.

The graphs in Figure 1 show the results. The top graph is for short queries and the bottom graph is for long queries. The x-axis is word error rate, and the systems from left to right are ordered as per Table 1. *I.e.* the leftmost point is CUHTK-S1, the next one is Dragon98-S1, ..., and the rightmost point is DERASRU-S1. The y-axis is non-interpolated average precision. Several interesting facts can be inferred from Figure 1:

Long queries are better than short queries. The average precision for human transcriptions is 0.5369 for long queries, whereas it is 0.4277 for short queries. This improvement in the quality of the queries is reflected in retrieval from all automatic transcriptions. For example, retrieval from DERASRU-S2 has an average precision of 0.3348 for short queries; this number rises to 0.4544 for long queries.

Loss of effectiveness is small for good transcriptions. Comparing the two solid lines—the horizontal one for human transcriptions and the other one for retrieval from various automatic transcriptions—we notice that the loss of effectiveness for reasonable automatic transcriptions is minimal. For both query sets, the effectiveness of retrieval from the best automatic transcription (CUHTK-S1) is almost the same as retrieval effectiveness for perfect text. The losses for other reasonable transcriptions (all but NIST-B2, DERASRU-S2, and DERASRU-S1) are all from 2–6%, which is minimal considering that these transcriptions have word error rates of up to 36%.

[1] In reality, the word error rate for human transcriptions is of course non-zero.

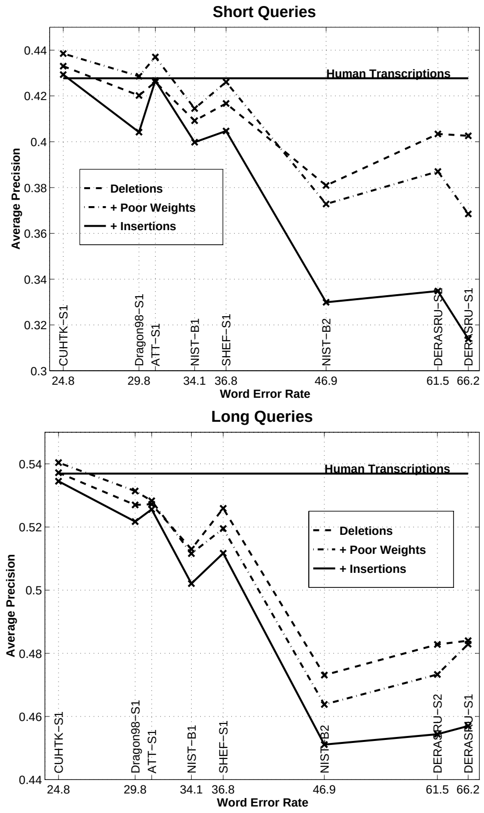

Figure 1: Losses due to different recognizer mistakes.

However, when recognition becomes noticeably poor, the retrieval effectiveness does fall about 22-27% for the short queries and about 15% for the long queries. *This observation strongly advocates the use of long queries whenever possible.* Use of long queries cuts the effectiveness difference between retrieval from human and automatic transcriptions.

Deletions matter. For good transcriptions, the differences in retrieval effectiveness for human and automatic transcriptions are little to begin with, and further break-up analysis for such small differences isn't expected to be very meaningful. But we do find that for the poor transcriptions, there is a noticeable loss in average precision due to word deletions (for both the short and the long queries).

Weight changes are less important. As we hypothesized earlier, the incremental loss due to poor term-weighting is relatively small. The main reason for this is the normalization of term frequencies done by the term weighting schemes used in modern IR systems, like the double-log normalized term frequency factor used by us. A term that occurs *tf* times in a document gets a weight of $1 + ln(1 + ln(tf))$ (ignoring document length effects). If a recognizer doesn't recognize the second occurrence of a word in a document, the loss is term weight is just 34% (its *tf*-weight become 1.0 instead of 1.5); whereas if a recognizer recognizes two out of three occurrences of a term, the loss in weight is just 12% (its *tf*-weight become 1.5 instead of 1.7).

Insertions matter for short queries. Further loss due to spurious words or insertions is quite noticeable for short queries, but it is relatively small for the long queries. Most of the loss for long queries is from deletions. This is evident in the bottom graph in Figure 1 from the large gap between the dashed-line (labeled "deletions") and the horizontal line. In the same graph, the incremental loss due to insertions (the solid line marked "+ Insertions") is not as large. This behavior is expected in our opinion. When queries are short they have few terms, and one or two spurious matches (the main effect of insertions) can dramatically promote the rank of a document. But once the queries have enough context (via more terms), there is more to match, and one or two spurious matches don't effect the ranking much since the overall ranking is governed by the entire context in the query.

5.2 Document Expansion

It is hard to separate insertions from spoken words, but we can reduce the degree of deletions via some kind of document expansion. If an automatic recognition can be enhanced with words that *could have been there* but didn't make it into the automatic transcription, then the resulting enhanced transcription should have fewer deletions and we should be able to cut our losses due to deletions. From a speech recognition perspective, an obvious way to bring new words into a document is via the use of alternative recognition hypotheses (for example by use of lattices or n-best transcriptions of a speech) [7]. Unfortunately we couldn't experiment with this type of document expansion since the lattices or the n-best transcriptions are not available to researchers.[2]

From an IR perspective, pseudo-feedback, which basically is nearest-neighbor based expansion, is an obvious way to bring related words into a text. We explore document expansion using Rocchio's method and study its effect on retrieval performance [19]. The main idea behind such document expansion is to, given a document, first find some documents that are related to the document at hand (its nearest neighbors), and then bring frequently occurring words from the related documents into this document. This process should be especially effective if the neighboring document are from a text corpus that is closely related to the speech at hand. Here are the steps involved in document expansion:

1. Select a collection of documents that will serve as the source of related documents. We use the North American News corpus available from LDC (LDC Catalog Number: LDC95T21, see `www.ldc.upenn.edu`) as the source of related documents. The main motivations behind using this collection are: 1) it is similar in nature to the speech collection at hand (both primarily contain American news), and 2) it contains print news from the same time period as the test data. Therefore we expect stories that are reported in the speech collection to also appear in this corpus. Since the test data is dated from June 1997 to January 1998, we used news dated from May 1997 to February 1998 (one month before and after) from the North American news corpus. From now on we will use *NA News* as a shorthand for this subset of the corpus.

2. Find documents related to a speech document. We do this by running the automatic transcription of the speech document as a query (*raw-tf*x*idf* weighted) on the NA News corpus and retrieving the ten most similar documents. In other words, we use the ten nearest neighbors of the speech document in this process. The automatic recognition of documents is weighted by *raw-tf*x*idf* (instead of using, say, a logarithmic or a

[2] Various groups have only given their one-best transcriptions to NIST.

Degree of Expansion	α in Rocchio			
	0.5	1.0	1.5	2.0
0%	0.4739	0.4656	0.4591	0.4556
10%	0.4898	0.4834	0.4756	0.4708
20%	0.4938	0.4866	0.4813	0.4763
50%	0.4990	0.4977	0.4898	0.4800
100%	0.4992	**0.4995**	0.4952	0.4892
200%	0.4943	0.4974	0.4923	0.4858

Table 2: Parameter exploration for NIST-B1 (short queries).

double-log *tf*-factor) when used as a query because we observed that nearest neighbors found using *raw-tf*x*idf* weighted documents yield the best expansion results. Space limitations don't allow us to detail those experiments here.

3. The speech transcriptions are then modified using Rocchio's formula.

$$\vec{D}_{new} = \alpha\vec{D}_{old} + \frac{\sum_{i=1}^{10} \vec{D}_i}{10}$$

where $\vec{D}_{old}$ is the initial document vector, $\vec{D}_i$ the the vector for the i-th related document, and $\vec{D}_{new}$ is the modified document vector. All documents are *dnb* weighted (see Table 1 in [25]). Optionally new words are added to the document. For term selection, the Rocchio weights for new words are multiplied by their *idf*, the terms are selected, and the *idf* is stripped from a selected term's final weight. Furthermore, to ensure that this document expansion process doesn't change the effective length of the document vectors, and change the results due to document length normalization effects, [24] we force the total weight for all terms in the new vector to be the same as the total weight of all terms in the initial document vector.

This approach is reminiscent of the use of single nearest neighbor clusters by Griffiths et al. in [6] and the use of five by nearest neighbors by Croft et al. in [3].

Various parameters are involved in this document expansion process. For example, how many nearest neighbors of a document should be used? We use 10. What should be α in Rocchio's process? We experimented with various values and picked the best. What should be the degree of document expansion? Once again, we tried different values and picked the best. A typical parameter tuning run is detailed in Table 2, which shows the retrieval effectiveness for various α values when the document is expanded by 0% of its original length (*i.e.* no expansion but the existing terms are reweighted), 10% of its original length (*i.e.* if the original document has 60 indexed terms, then we add 6 new terms to the document), ..., 200% of its original length. We found that for short queries, $\alpha = 1.0$ with 100% expansion worked the best for all transcriptions. For long queries $\alpha = 1.5$ *or* 2.0 with 50–100% expansion was the best. However, for these queries, the difference in retrieval performance with $\alpha = 1.0$ and 100% expansion (the best parameter setting for short queries) was under 1% for most of the cases so we decided to use $\alpha = 1.0$ and 100% expansion in all our experiments.

Results from Document Expansion

We ran both the query sets (short and long) on the modified and the original documents and measured average precision.

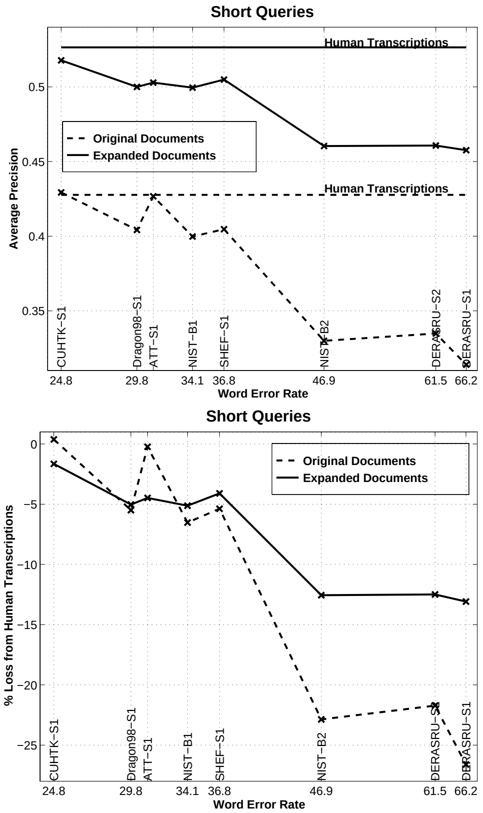

Figure 2: Document expansion for short queries.

The results for the short queries are shown in Figure 2, and those for the long queries are shown in Figure 3. For a fair comparison, we cannot compare the results from expanded automatic transcriptions to unexpanded human transcriptions. It is possible that document expansion is generally helpful for this collection and it does not hold any added advantage for speech transcriptions. Therefore, the baseline for comparing our expanded speech transcriptions results is the result from the *expanded* human transcriptions.

Many interesting facts can be observed from Figures 2 and 3. Let's analyze the results for the short queries first. The top plot in Figure 2 plots the average precision on the y-axis against the word error rates for various transcripts on the x-axis. The dashed lines are for unexpanded (original) documents, and are the same as the two solid lines in the top graph in Figure 1. The solid lines are for the expanded documents. The horizontal lines correspond to the average precision for retrieval from human transcripts (dashed line) and expanded human transcripts (solid line).

First of all, we observe that document expansion dramatically improves the average precision for short queries for *all* transcriptions. We expected document expansion to improve average precision for automatic transcriptions, but the 23% improvement for perfect text (the average precision jumps from 0.4277 to 0.5265) is quite unexpected. Previous studies have shown modest gains when spreading activation was used with five nearest neighbors of a document [3]. Whether this effect will hold when applied to large text collections (like TREC [28]) is still unclear. We will test document expansion on large collections in near future. Figure 2 shows that, like for the human transcripts, document expansion also improves the retrieval effectiveness

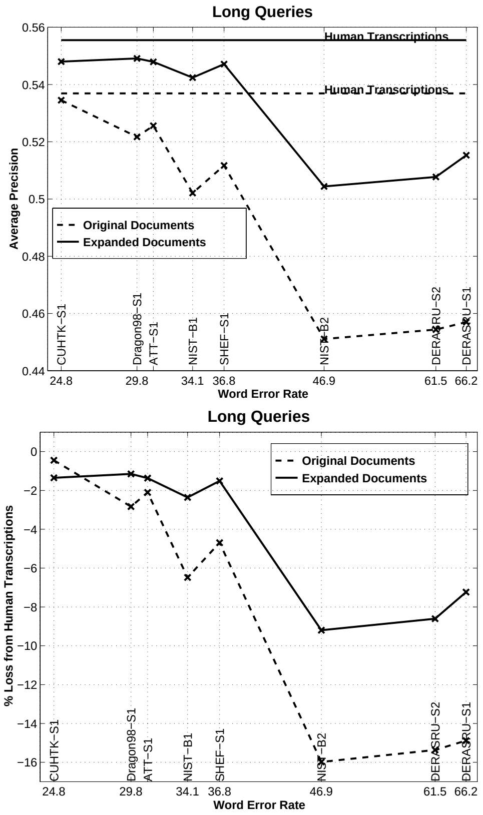

Figure 3: Document expansion for long queries.

for automatic transcripts by a large margin. For example, the retrieval effectiveness for DERASRU-S1 transcripts (the rightmost point on the graphs) jumps from 0.3139 to 0.4576, a remarkable improvement of over 46% (0.3139 being the baseline).

As we mentioned above, it might be the case that for this test collection document expansion is beneficial in general, and it doesn't hold any special advantage for automatic speech transcripts. However, the bottom graph in Figure 2 shows that this is not the case, and document expansion indeed is more useful when the text is erroneous. The dashed line on the bottom graph shows the %-loss in average precision when retrieval is done from (unexpanded) automatic transcriptions instead of (unexpanded) human transcriptions. This line has the same shape as the dashed line on the top graph since it is essentially the same curve on a different scale (0 to −100%, the human transcriptions being the 0% mark). And we notice that the loss for CUHTK-S1 (the leftmost point) is close to 0% whereas it is 27% for DERASRU-S1 (the rightmost point). The solid line on the bottom plot shows the %-loss for various transcripts for expanded documents. The baseline for this curve is higher; it corresponds to the solid horizontal line on the top graph. We see that document expansion indeed benefits the poor transcriptions much more that it benefits the human or the better automatic transcriptions. For poor transcriptions, the gap in retrieval effectiveness reduces from 23% to about 13% for NIST-B2, from 22% to about 13% for DERASRU-S2, and from about 27% to about 13% for DERASRU-S1. All these loss reductions are quite significant.

In summary, document expansion is more useful for automatic speech transcripts than it is for human transcriptions. Automatic recognitions that are relatively poor need the most help during retrieval. Document expansion is helping exactly these transcriptions, and quite noticeably. It is encouraging that even with word error rates as high as 65%, the retrieval effectiveness drops just 12-13% post document expansion. This drop would have been 22-27% without expansion. This effect will be further reinforced by our following observations for long queries.

Now studying similar graphs for the long queries in Figure 3, we observe that document expansion is once again beneficial for all transcripts, though not quite as much as it was for the short queries. For example, for human transcriptions, document expansion yields an improvement of 23% for short queries (over no document expansion). This improvement is just 3.5% for the long queries (which is very much inline with the improvements suggested by earlier studies that use nearest neighbors [3]). We believe this happens because when queries are short, they stand to gain from document enrichment done by expansion. On the other hand, when queries are already rich in content, like the long queries, the incremental benefits from enriched documents are minimal. This in itself is an interesting result.

More interestingly, we observe that when documents are poor in content, like the automatic transcripts with lots of errors, document expansion is still beneficial despite the richness of the queries. Even with long queries, we see a marked improvement in retrieval effectiveness from poor transcription. The average precision improves about 12-13% for the three transcriptions with high word error rates (as opposed to just 3-5% for perfect text or the better transcriptions), and the gap is reduced to about 8% instead of the original 16%. Once again we find that document expansion is helping where help is needed most, that is, for poor automatic transcriptions.

5.3 Further Analysis

Document expansion has two effects on a document vector. First, Rocchio's method *reweights* the terms that already exist in the document. Second, it brings new words in the document. We now study how these two effects change the retrieval process. Specifically, we examine how much of the improvement from document expansion is due to reweighting of the existing terms, and how much of it is due to addition of new terms.

Figure 4 shows the effects of reweighting only and of adding new terms. The top graph is for short queries and the bottom graph is for long queries. Once again the x-axis is WER and the y-axis is average precision. The horizontal lines are for human transcriptions and the other lines are for various automatic transcriptions. For the top graph in Figure 4 the dashed lines (original documents or no expansion) and the solid lines (full expansion) are the same as in the top graphs of Figure 2; and in the bottom graph they correspond to the top graph of Figure 3. The additional lines in Figure 4 (drawn with dashes and dots) show the effect of reweighting only.

The horizontal lines in Figure 4 show that for perfect text (human transcriptions), majority of the improvements from document expansion are due to reweighting of the existing terms. For short queries, just by reweighting, the average precision jumps 17% from 0.4277 to 0.5017. This gain is about 4% for long queries. This, we believe, is due to the redistribution of weights that occurs for the terms present in a document. More specifically, all terms that appear equally often in a document get equal weights before reweighting,

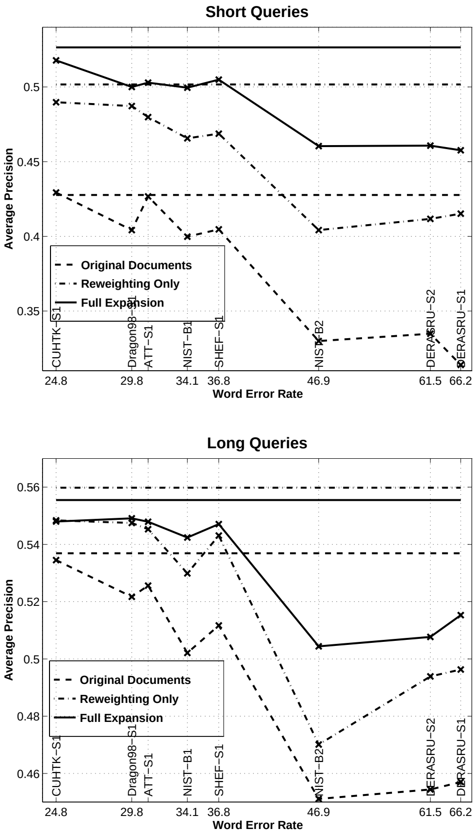

Figure 4: Effect of reweighting and new words.

but the presence of some of those terms is supported by the nearest neighbors of the document, whereas the presence of other terms is not supported. After reweighting, this results in a higher weight for terms whose presence in the document is supported by nearest neighbors and vice-versa. For example, if a document is about automatic speech recognition, and the words `Markov` and `spectrum` both occur just once in it, the nearest neighbors for this document might support the presence of the word `Markov` more strongly than they would support the presence of `spectrum`, yielding a higher weight for the word `Markov` in the reweighted document as compared to the weight of `spectrum`. Put another way, there are some "signal" words in a document and there are some "noise" words. Document reweighting based on nearest neighbors emphasizes the signal words and it deemphasizes the noise words yielding a better overall term weight assignment for the document.

Adding new terms to the documents adds another 6% for the short queries yielding a final average precision of 0.5265 (a 23% improvement over 0.4277). We believe that adding new terms to documents is having the same effect as using long queries. Our belief is supported by the fact that when queries are indeed long, we don't see any improvement due to adding new terms to documents. Adding new terms to documents actually results in a small loss when queries are long (see horizontal lines in bottom graph of Figure 4).

The effects of reweighting the existing terms is similar for all automatic speech transcriptions. *I.e.,* we see a large improvement in retrieval effectiveness when certain words, which the document is truly about, get higher weights. But we notice that adding new terms to the automatic transcriptions is more useful than it is for human transcriptions. Overall, adding new terms never hurts for speech transcriptions, and is marginally to noticeably useful. This result is important since it shows that addition of new terms helps automatic transcripts (it helps noticeably when the transcripts are poor), even though it might hurt the perfect transcriptions (as it does for the long queries).

6 Expansion from Unrelated Corpus

Above results show that when we have a text corpus which is reasonably close to the speech corpus in content type, then document expansion is truly beneficial for speech retrieval. But what happens when we don't have such a corpus available to us? To study the effect of document expansion when it is done from a corpus that does not closely related to the speech at hand, we used TREC disks 1–5 as the corpus for document expansion. This is a large corpus of about 5.2 Giga-bytes containing 1,634,976 documents from various sources (news and non-news) [28]. The news material in this corpus is from the years 1987–1994 and has little overlap with the news topics covered in our speech corpus (which is dated from June 1997 to January 1998).

Space limitation prohibit us from presenting graphs for this set of experiments, but the main highlight of these experiments is that *document expansion from the TREC corpus is not as beneficial as it is from the closely related NA News corpus.* For example, when documents were expanded from NA News, the average precision for retrieval from perfect text using short queries jumped 23% from 0.4277 to 0.5265; but when documents are expanded from TREC, this increase is just 12% (from 0.4277 to 0.4828). Document expansion actually hurts retrieval effectiveness for long queries. We now lose about 0.7% in average precision as opposed to a 3.5% gain for NA News.

More importantly, we find that the performance gap between the human and the ASR transcripts is not reduced as significantly as it did when documents were expanded from NA News. E.g., for the short queries and NIST-B2 transcripts, the gap now reduces from 23% to just 20% (whereas it reduced to 12% for NA News). This situation is worse for long queries for which this gap doesn't change much for most of the transcripts. Document expansion is still useful for all transcripts when queries are short, though not as much; whereas it has almost no effect when queries are long.

Further analysis shows that most of the effect of document expansion from TREC is due to reweighting of the existing terms. For the long queries, bringing in new terms doesn't help much (either helps or hurts about 1%). However, for the short queries, it still does help (about 2–7% depending upon the transcripts), though much less than expansion from NA News. Changing the parameter values for document expansion from TREC doesn't change the results much. Overall, it appears that the effectiveness for document expansion for better retrieval is largely dependent upon having a text collection for document expansion that is closely related to the speech at hand.

Yet another possible source for document expansion is the speech corpus itself. In this scenario, the nearest neighbors of a speech document will be other speech documents. This would completely eliminate the need for an external text collection for doing document expansion. Space limitations don't allow us to report our experiments here, but document expansion from this corpus itself isn't very effective. The main problem is the small size of this corpus. Most documents don't have any related documents in the corpus.

7 Conclusions

Our main results are as follows. First, document expansion from a text collection closely related to the speech at hand yields substantial benefits for speech retrieval, reducing the performance gap between retrieval from perfect text and from automatic speech transcriptions. Retrieval from reasonable speech transcriptions is competitive with retrieval from perfect text. Document expansion helps where help is needed most, namely, for poor automatic transcriptions. On the other hand, expansion from unrelated corpora is not nearly as beneficial.

Second, retrieval using long queries is more robust against speech recognition errors. Term insertion by a speech recognizer is not a significant problem if the queries have enough context (long queries). Furthermore, using modern term weighting schemes, the losses incurred due to improper *tf*-values for terms that were actually spoken and were also recognized, are minimal.

This study also raises a number of interesting further questions. For short queries, the improvements obtained by expanding perfect text are unexpected. Traditional wisdom (as in use of document clusters or other devices) says that this should not be the case. This effect should be studied on a larger text collection like TREC. Also the dependence of document expansion on the availability of a closely related text collection should be studied further. It would be desirable to get similar results when a closely related text collection is not available.

Acknowledgments

We are very grateful to all the SDR track participants who have provided their transcriptions to NIST; and to NIST for making them available to us.

References

[1] C.W. Cleverdon and J. Mills. The testing of index language devices. *Aslib Proceedings*, 15(4):106–130, 1963.

[2] P. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation by semantic networks. *Information Processing and Management*, 23(4):266–268, 1987.

[3] W.B. Croft, T.J. Lucia, J. Cringean, and P. Willett. Retrieving documents by plausible inference: An experimental study. *Information Processing and Management*, 25(6):599–614, 1989.

[4] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic indexing. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] Edward Fox, Gary Nunn, and Whay Lee. Coefficients for combining concept classes in a collection. In *Proceedings of SIGIR'88*, pages 291–307, 1988.

[6] A. Griffiths, H.C. Luckhurst, and P. Willett. Using inter-document similarity in document retrieval systems. *Journal of the American Society for Information Science*, 37:3–11, 1986.

[7] A.G. Hauptmann, R.E. Jones, K. Seymore, S.T. Slattery, M.J. Witbrock, and M.A. Siegler. Experiments in information retrieval from spoken documents. In *Proceedings of the Broadcast News Transcription and Understanding Worshop*, pages 175–181, 1998.

[8] W.J. Hutchins. The concept of "aboutness" in subject indexing. *Aslib Proceedings*, 30:172–181, 1978.

[9] N. Jardine and C.J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971.

[10] Y. Jing and W.B. Croft. An association thesaururs for information retrieval. In *RIAO 94 Conference Proceedings*, pages 146–160, October 1994.

[11] G.J.F. Jones, J.T. Foote, K. Sparck Jones, and S.J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of SIGIR'96*, pages 30–38. ACM, New York, August 1996.

[12] K. Sparck Jones. *Automatic Keyword Classification for Information Retrieval.* Butterworths, London, 1971.

[13] T. Joyce and R.M. Needham. The thesaurus approach to information retrieval. *American Documentation*, 9:192–197, 1958.

[14] K.L. Kwok. The use of titles and cited titles as document representations for automatic classification. *Information Processing and Management*, 11(8-12):201–206, 1975.

[15] K. Ng and V. Zue. Subword unit representations for spoken document retrieval. In *Eurospeech 97*, pages 1607–1610, 1997.

[16] Y. Qiu and H.P. Frei. Concept based query expansion. In *Proceedings of SIGIR'93*, pages 160–169, June 1993.

[17] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[18] S.E. Robertson and S. Walker. Some simple effective approximations to the 2–poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR'94*, pages 232–241. Springer-Verlag, New York, July 1994.

[19] J.J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System—Experiments in Automatic Document Processing*, pages 313–323, Englewood Cliffs, NJ, 1971. Prentice Hall, Inc.

[20] R.C. Rose. Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 41(1):45–60, 1991.

[21] Gerard Salton. Associative document retrieval techniques using bibliographic information. *Journal of the American Society for Information Science*, 10(4):440–457, October 1963.

[22] Gerard Salton. Cluster search strategies and the optimization of retrieval effectiveness. In Gerard Salton, editor, *The SMART Retrieval System—Experiments in Automatic Document Processing*, pages 223–242, Englewood Cliffs, NJ, 1971. Prentice Hall, Inc.

[23] Gerard Salton and M.E. Lesk. Computer evaluation of indexing and text processing. *Journal of the Association for Computing Machinery*, 15:8–36, 1968.

[24] Amit Singhal. *Term Weighting Revisited.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, NY, January 1997.

[25] Amit Singhal, John Choi, Donald Hindle, David Lewis, and Fernando Pereira. AT&T at TREC-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).

[26] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *Journal of the American Society for Information Science*, 45(1):50–58, January 1994.

[27] Ellen Voorhees. *The effectiveness and efficiency of agglomerative hierarchical clustering in document retrieval.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, NY, January 1986.

[28] E.M. Voorhees and D.K. Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).

[29] M. Wechsler, E. Munteanu, and P. Schauble. New techniques for open vocabulary spoken document retrieval. In *Proceedings of SIGIR'98*, pages 20–27. ACM Press, August 1998.

[30] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–598, 1988.

[31] M.J. Witbrock and A.G. Hauptmann. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proceedings of DL'97*, pages 30–35. ACM, 1997.