Crestani, F. (2002) *Spoken query processing for interactive information retrieval.* Data and Knowledge Engineering, 41 (1). pp. 105-124. ISSN 0169-023X

Spoken Query Processing for Interactive Information Retrieval

Fabio Crestani

Department of Computer and Information Sciences
University of Strathclyde
26 Richmond Street
Glasgow G1 1XH, Scotland, UK

Abstract

It has long been recognised that interactivity improves the effectiveness of Information Retrieval systems. Speech is the most natural and interactive medium of communication and recent progress in speech recognition is making it possible to build systems that interact with the user via speech. However, given the typical length of queries submitted to Information Retrieval systems, it is easy to imagine that the effects of word recognition errors in spoken queries must be severely destructive on the system's effectiveness. The experimental work reported in this paper shows that the use of classical Information Retrieval techniques for spoken query processing is robust to considerably high levels of word recognition errors, in particular for long queries. Moreover, in the case of short queries, both standard relevance feedback and pseudo relevance feedback can be effectively employed to improve the effectiveness of spoken query processing.

Key words:
PACS: Information Retrieval, Spoken Query Processing, Evaluation.

1 Introduction

The widespread access to Internet and the high bandwidth often available to users makes them assume that there is nothing easier than remotely connecting to and searching a large information repository. Nevertheless, a very large

Email address: F.Crestani@cs.strath.ac.uk (Fabio Crestani).
URL: http://www.cs.strath.ac.uk/~fabioc/ (Fabio Crestani).

part of the world population does not and will not have for a long time access to computers or Internet. There are cases in which the only available (or the most convenient) communication mean is a telephone or mobile phone. In addition, there are situations and users for whom the availability of computer screen and keyboard is not useful. Blind or partially-sighted users (e.g. users who have problems due to disabilities, protective clothing or working environment) may only have access to information if it is accessible via voice and/or sound. In all these cases, if we want users to take advantage of the large amount of information stored in digital repositories, it is necessary to enable access to them solely via voice. This implies the design and implementation of systems capable not only of understanding the user's spoken request, finding the required information and presenting it as speech, but also capable of interacting with the user in order to better understand the user information need, whenever this is not clear enough to proceed effectively with the searching.

Recent progress in speech recognition is making it possible to build systems that interact via voice with the user. Nevertheless, speech recognition systems are, and for long time still will be, affected by recognition errors. The effects of word recognition errors (WRE) in spoken documents on the performance of Information Retrieval (IR) systems have been well studied and well documented in recent IR literature. A large part of the research in this direction has been promoted by the Spoken Document Retrieval (SDR) track of TREC (see for example [1]). Experimental evidence has brought to the conclusion that for long documents and for reasonable levels of average Word Error Rates (WER), the presence of errors in document transcripts does not constitute a serious problem. In a long document, where terms important to the characterisation of the document content are often repeated several times, the probability that a term will always be misrecognised is quite low and there is a good chance that a term will be correctly recognised at least once. Variations of classical IR weighting schemes (for example giving a lesser importance to the within-document term frequency) have been proposed that are able to cope with reasonable levels of WER [2], however these solutions were not found so effective for short documents, for which SDR has not found an effective solution yet.

In contrast with SDR, very little work has been carried out in the area of Spoken Query Processing (SQP), that is the use of spoken queries to retrieve textual or spoken documents. In particular, very little research work has been devoted to studying the effects of WRE in SQP. A spoken query can be considered similar to a very short document and high levels of WER may have devastating effects on the performance of an IR system. In a query, as in a short document, the misrecognition of a term may cause it to disappear completely from the query representation and, as a consequence, a large set of potentially relevant documents indexed using that term will not be retrieved.

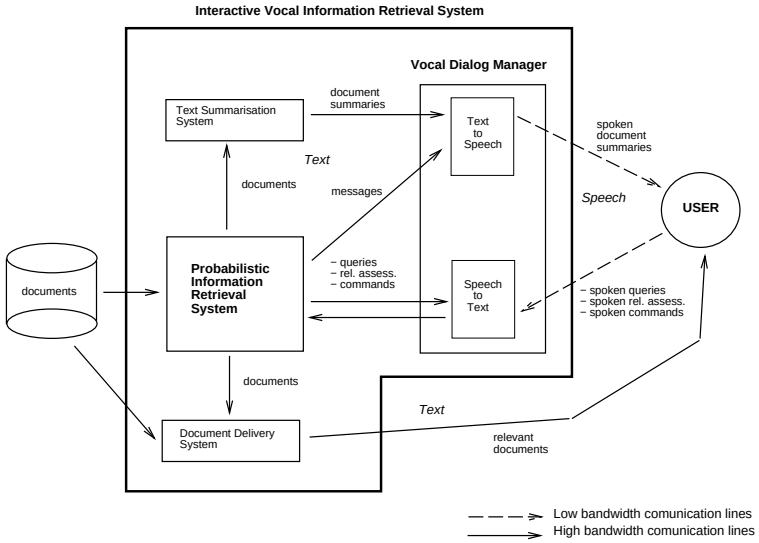

Figure 1. Schematic view of the IVIRS prototype

In this paper we present the results of an experimental study of the effects of WRE in spoken queries on the effectiveness of an IR system. The paper is structured as follows. Section 2 explain the background of this work. Section 3 highlights the differences between SQP and SDR. This section also gives the motivations of the work reported here. Section 4 presents the experimental environment of the study: the test collection, the queries, and the IR system used. The effects of WRE in spoken queries on the effectiveness of the IR system are presented and discussed in section 5. Section 6 reports on the use of relevance feedback as a way to improve the effectiveness of the SQP task. The limitations of the study are reported in 7. Section 8 briefly introduces other techniques for improving SQP that are currently being investigated. The conclusions of the study are reported in section 9.

2 The SIRE Project

The background of the work reported in this paper is related to the *Sonification of an Information Retrieval Environment* (SIRE) project. The main objective of the project is to enable a user to interact (i.e. submit queries, commands, relevance assessments, and receive summaries of retrieved documents) with a probabilistic IR system over a low bandwidth communication line, like for example a land telephone line or a mobile phone line. A first prototype system, called *Interactive Vocal Information Retrieval System* (IVIRS), is currently being developed [3]. An outline of the system specification of the prototype is reported in figure 1.

IVIRS works in the following way. A user connects to the system using a

telephone. After the system has recognized the user by means of a username and a password, the user submits a spoken query to the system. The Vocal Dialog Manager (VDM) interacts with the user to identify the exact part of the spoken dialogue that constitutes the query. The query is then translated into text and fed to the probabilistic IR system (PIRS). Additional information regarding the confidence of the speech recognizers is also fed to the PIRS. The PIRS searches the textual archive and produces a ranked list of documents, and a threshold is used to find the set of documents regarded as very likely to be relevant (this feature can be set in the most appropriate way by the user). The user is informed on the number of documents found to be relevant and can submit a new query or ask to inspect the documents found. Documents in the ranked list are passed to the Document Summarization System that produces a short representation of each document that is then read to the user over the telephone using the Text-to-Speech module of the VDM. The user can wait until a new document is read, ask to skip the document, mark it as relevant or stop the process altogether. Marked documents are stored in a retrieved set and the user can proceed with a new query if he wishes so. A document marked as relevant can also be used to refine the initial query and find additional relevant documents by feeding it back to the PIRS. This interactive relevance feedback process can go on until the user is satisfied with the retrieved set of documents. Finally, the user can ask the documents in the retrieved set to be read in their entirety or delivered by email or ftp via the Document Delivery System.

For the prototype implementation of IVIRS a "divide and conquer" approach has been followed, consisting of dividing the implementation and experimentation of IVIRS in the parallel implementation and experimentation of its different components. Work carried out so far has concentrated on implementing and experimenting with the Document Summarization System [4], the Text-to-Speech and Speech-to-Text modules of the VDM [5], and the Document Delivery System.

3 Spoken Query Processing for Information Access

One of the underlying assumptions of the design of IVIRS is that the spoken queries could be recognized by the VDM with a level of correctness as to enable their effective use by the PIRS. However, this assumption cannot be supported by past research. In fact, while a number of studies have been devoted to studying the effects of WRE in SDR, much less research has addressed the effects of WRE in SQP.

It has to be recognized that SQP poses a number of additional challenges compared with SDR. The most important ones are:

(1) query processing needs to be performed on-line and "almost" real time, while spoken document recognition and indexing can be performed off-line;
(2) queries are usually much shorter than documents and WRE may have more serious effects on them;
(3) we may have very little training data on the voice of each user and we may have a large number of different users in different acoustic conditions.

In SDR, spoken documents are almost always processed off-line using large vocabulary speech recognition (SR) techniques. This is due to the computationally intensive nature of the SR process. The time required by a SR system to process a spoken document depends on the length of the document, on the system, and on the machine the system is operating. It is not unusual to require 200 time units to process one time unit of speech [1]. This does not constitute a serious problem in SDR, where spoken documents are processed off-line to produce transcripts and transcripts are processed off-line to produce IR indexes. SQP, on the other hand, requires that queries are processed on-line, at the time they are submitted by the user. A spoken query needs to be speech-processed and a transcript needs to be produced at the time the query is submitted. In addition the transcript needs to be indexed and matched against the IR document indexes on-line, as it is done in any text-based IR application. It has been observed that user satisfaction with an IR system is dependent also upon the time the user spends waiting for the system to process the query and display the results [6], therefore it is advisable that this time is kept short, in the order of few seconds. Although queries are usually much shorter than documents and therefore the time necessary to speech-process them is shorter, this requirement should still be kept in mind when designing SQP systems. The second issue is related to the effectiveness of SQP. It is a well known fact in textual IR that short queries are less effective than long queries in finding relevant documents [7]. This is in large part due to the so called "term mismatch problem". The causes of this problem are related to the fact that users of IR systems often use different terms to describe the concepts in their queries than the authors use to describe the same concepts in their documents. It has been observed that two people use the same term to describe the same concept in less than 20% of the cases [8]. It has also been observed that this problem is more severe for short casual queries than for long elaborate ones since, as queries get longer, there is a higher chance of some important term co-occurring in the query and the relevant documents [9]. The term mismatch problem does not have only the effect of hindering the retrieval of relevant documents, it has also the effect of producing bad rankings of retrieved documents. The term mismatch problem becomes more severe when it is combined with the "term misrecognition" problem. In fact, the misrecognition of a spoken query term may cause the term to disappear completely from the query representation or, even worse, to be replaced by a different term. Because of the term misrecognition problem a large set of potentially relevant

documents indexed using that term may not be retrieved. We shall see in the experimental analysis reported in the remainder of this paper how severe this problem really is.

The third issue is related to the difficulty of correct recognition of terms in a spoken query. SR systems usually rely on some training data to fine-tune the SR system on the data to be recognized. The training data is usually very similar to the data to be recognized, so that some of the parameters of the SR process can be tuned on the data. In SDR this testing and tuning of the system is almost always performed. However, this may not be possible in SQP, since it may be the first time the user submits a query (so no previous data are available on the user voice and vocabulary to be used to fine-tune the system) or the acoustic conditions may be different (for example, the user may be submitting the query in a different acoustic environment or using a different microphone/telephone). The lack of training data may cause the performance of the SR process to be poor and spoken queries may have exceptionally high WER.

The effects of the above issues on the effectiveness of an IR system engaged in SQP have not been fully studied yet. The work reported in this paper tries to partially amend this lack.

4 Experimenting with Spoken Query Processing

It is the author's view that the best way to assess the limitations of current technologies for SQP is to conduct an experimental analysis using state of art techniques and tools in a setting as close as possible to a real world application.

In order to experiment the effects of WRE in SQP a suitable test environment needs to be devised. Classical IR evaluation methodology [10] suggests that we use the following:

(1) a collection of textual document;
(2) a set of spoken queries with associated relevance assessments recognised at different levels of WER;
(3) an IR system;
(4) some way of measuring the IR system effectiveness.

In the following we describe these components of our experimental environment.

Data set:	WSJ 1990-92
Number of documents	74.520
Size of collection in Mb	247
Unique terms in documents	123.852
Average document length	550
Average document length (unique terms)	180

Table 1
Characteristics of the Wall Street Journal 1990-92 document collection.

4.1 The Test Collection

The collection we used is a subset of the collection generated for TREC (see for example [11]). The collection is made of the full text of articles of the Wall Street Journal (years 1990-92). Some of the characteristics of this test collection are reported in table 1.

We used the full text of documents after the SGML tags were removed. No use of the HL (headline) or LP (leading paragraph) tags was made, as opposed to most system participating in TREC, and the text of all sections of the document was considered indistinctively.

4.2 The Spoken Queries

A set of 35 queries (topics 101-135 of TREC) with the corresponding lists of relevant documents was used. These queries were originally in textual form and were quite long, however some of the fields of the query were not used in the experiments reported in this paper. In fact, the only field used were title, description, and concepts, but considering the text in them indistinctively. This makes the queries short enough to be a somewhat more realistic examples of "real" user queries. Some of the characteristics of these queries are reported in table 2.

Since the original queries were in textual form, it was necessary to reproduce then in spoken form and have then recognised by a SR system. This work was carried by Jim Barnett and Stephen Anderson of Dragon Systems Inc. [12]. Barnett and Anderson had one single (male) speaker dictate the queries. The spoken queries were then recognised by Dragon's research LVCSR system, a SR system that has a 20,000 words vocabulary and a bigram language model trained on the Wall Street Journal data.

The bigram language model grows with the vocabulary, and even a reduced

Data set:	Topics 101-135
Number of queries	35
Unique terms in queries	3.304
Average query length (with stopterms)	58
Average query length (without stopterms)	35
Median query length (without stopterms)	28
Average number of relevant documents per query	30

Table 2
Characteristics of topics 101-135 of TREC.

Query sets	27	28	29	34	35	47	51	75
Avg. num. substitutions %	18.8	19.1	20.0	22.7	24.2	31.5	35.5	49.8
Avg. num. deletions %	2.6	2.6	2.6	2.6	2.6	3.0	4.2	2.9
Avg. num. insertions %	6.0	6.0	6.6	8.4	7.8	12.4	11.3	21.8
Avg. num. errors %	27.4	27.7	29.2	33.6	34.6	46.8	51.0	74.5
Avg. num. sentence errors %	39.1	40.0	40.4	42.2	47.0	51.3	56.7	66.5

Table 3
Characteristics of the different query sets.

vocabulary of 20,000 words can lead to large search spaces respresented by the word lattice. Nevertheless the uneven distribution of probabilities among different paths can help. A complexity reduction technique called beam search, consisting of neglecting states whose accumulated score is lower than the best one minus a given threshold, if often used to limit the search. In this way, computation needed to expand bad nodes is avoided.

By altering the width of the beam search, a number of sets of transcripts at different levels of WER were generated. The beam width was chosen as the major parameter to alter because it was believed that this yields relatively realistic recognition errors. The standard error characteristics of these sets of transcripts are reported in table 3. We shall refer to these sets as query sets. More details on the process used to generate these different sets of transcripts are reported in [12].

In table 3, notice that each set has been denoted with a name referring to the approximate average WER of that set (i.e. the "Avg. num. errors %"). The set identified by 0 (not reported in the table, but reported later on as a reference line) is the perfect transcript, which contains no errors.

4.3 The Information Retrieval System

We used a modified version of an experimental IR toolkit developed at Glasgow University by Mark Sanderson [13]. The system is a collection of small independent modules each conducting one part of the indexing, retrieval and evaluation tasks required for classic IR experimentation. The modules are linked in a pipeline architecture communicating through a common token based language. The system is implemented on the UNIX operating system which, with its scripting and pre-emptive multi-tasking is eminently suitable for supporting such a modular architecture. This same system was used by the author for submissions to TREC (see for example [14]), and it was chosen as the IR platform for the experiments reported in this paper because it implements a model based on the classical $tf - idf$ weighting schema.

The *idf* (inverse document frequency) formula used by the system is:

$$idf(t_i) = -\log \frac{n_i}{N}$$

where n_i is the number of documents in which the term t_i occurs, and N is the total number of documents in the collection.

The tf (term frequency) is defined as:

$$tf(i,j) = \frac{\log(freq_{i,j} + 1)}{\log(length_j)}$$

where $freq_{i,j}$ is the frequency of term t_i in document d_j, and $length_j$ is the number of unique terms in document d_j.

The RSV of a document with respect to a query is evaluated by applying a similarity measure (the dot product) to the document and query representations obtained using the $tf - idf$ weighting schema. In other words, the score for each document is calculated by summing the $tf - idf$ weights of all query terms found in the document:

$$RSV(d_j, q) = \sum_{t_i \in q} idf(t_i) \cdot tf(i,j)$$

In the IR literature there exist many variations of this formula depending on the way the tf and *idf* weights are computed [15]. We chose this one because it is arguably the most standard scheme. Other weighting schemes may prove to be more or less effective.

A few state-of-the-art IR systems enable to perform Relevance Feedback (RF). *Standard RF* is a technique that enables a user to interactively express his information requirement by modifying his original query formulation with further information [16]. This additional information is provided by explicitly confirming the relevance of some indicating documents retrieved by the system. Obviously the user cannot mark documents as relevant until some are retrieved, so the first search has to be initiated by a query and the initial query specification has to be good enough to pick out some relevant documents from the collection. It is sufficient that at least one document in the list of retrieved documents matches, or come close to match, the user's interest, to initiate the RF process. The user can mark the document(s) as relevant and starts the RF process. If RF performs well the next list should be closer to the user's requirement and contain more relevant documents than the initial list. A different form of relevance feedback is *Pseudo RF*. In this method, rather than relying on the user to explicitly choose some relevant documents, the system assumes that its top-ranked documents are relevant, and uses these documents in the RF algorithm. This procedure has been found to be highly effective in some cases, in particular in those in which the original query is long and precise [16].

Sanderson's original system did not provide RF, but it was not too difficult to add a module implementing it. Among the many algorithms for RF, Probabilistic RF was chosen and implemented. Briefly, Probabilistic RF consists of adding new terms to the original query. The terms added are chosen by taking the first k terms in a list where all the terms present in relevant documents are ranked according to the following relevance weighting function [17]:

$$rw(t_i) = r_i \cdot \log \frac{(r_i + 0.5) \cdot (N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5) \cdot (n_i - r_i + 0.5)}$$

where: N is the number of documents in the collection, n_i is the number of documents with at lest one occurrence of term t_i, R is the number of relevant documents used in the RF, and r_i is the number of relevant documents in R with at least one occurrence of term t_i. The score for each document is then calculated using the following formula that uses the relevance weight $rw(t_i)$ of the term t_i instead of $idf(t_i)$.

$$RSV_{RF}(d_j, q) = \sum_{t_i \in q} rw(t_i) \cdot tf(i, j)$$

Essentially, Probabilistic RF compares the frequency of occurrence of a term in the documents marked as relevant with its frequency of occurrence in the whole document collection. If a term occurs more frequently in the documents marked as relevant than in the whole document collection it is assigned a higher $rw(t_i)$ weight. Then, there are two ways of choosing the terms to add

to the query: (1) adding terms whose weight is over a predefined threshold, or (2) adding a fix number of terms, for example the k terms with the highest $rw(t_i)$ weight. In the experiments reported in this paper we used the second technique. After a few tests, the number of terms to be added to the original query was set to 10.

4.4 Effectiveness Measures

The main IR effectiveness measures are Recall and Precision. *Recall* is defined as the portion of all the relevant documents in the collection that has been retrieved. *Precision* is the portion of retrieved documents that is relevant to the query. Once documents are ranked in response to a query using the RSV, precision and recall can be easily evaluated. These values are displayed in tables or graphs in which precision is reported for standard levels of recall (from 0.1 to 1.0 with 0.1 increments). We should remember that, experimentally, these measures have proved to be related in such a way that high precision brings low recall and vice versa.

Another useful effectiveness measure for IR is *average precision*, defined as the average of the 11 precision values of the different levels of recall.

In order to give a measure of the effects on the effectiveness of the IR system of different WER in the spoken queries, a number of retrieval runs were carried out with the different query sets and precision and recall values were evaluated. The results reported in the following graphs are averaged over the entire sets of 35 queries.

5 Word Recognition Errors and Effectiveness of Spoken Query Processing

This section reports some of the results of the experimental analysis of the effects of WRE in SQP. Not all the results of the experiments carried out are presented; a more complete report of this study can be found in [18].

5.1 Effects of Word Recognition Errors on Spoken Query Processing

The first experimental analysis was directed towards studying the effects of different WERs in spoken queries on the effectiveness of an IR system using a standard IR model. The parameters configuration most commonly used in

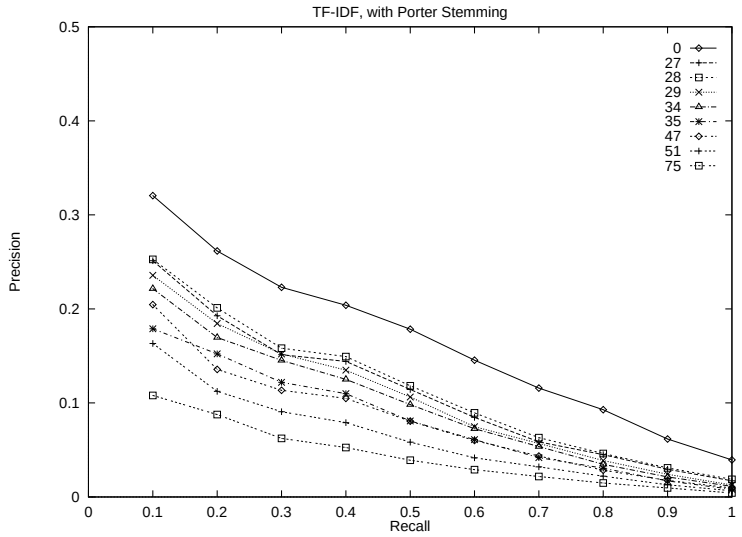

Figure 2. Results using the $tf - idf$ weighting scheme with stemming.

textual IR employs the $tf - idf$ weighting scheme on terms extracted from documents and queries. Extracted terms are first compared with a stoplist, i.e. a list of non content-bearing terms that can be removed from the IR indexes without loss of effectiveness. Terms appearing in the stoplist are removed, and the remaining terms are subject to a stemming and conflation process, in order to further reduce the dimensionality of the term space and to avoid a high incidence of the term mismatch problem due to trivial word variations, like for example singular/plural forms, verb tenses, and so on. In the experiments reported here a standard stoplist [19] and the stemming and conflation algorithm commonly known as "Porter algorithm" were used [20].

Figure 2 depicts the effects of different WERs in queries on the effectiveness of the IR system using the above standard configuration. Naturally, it can be noted that the best results are obtained for the perfect transcript (the transcript 0), and there is a degradation in effectiveness that can be attributed to WRE. Higher WERs cause lower effectiveness. An attentive reader can notice that the reference effectiveness (the one obtained with the perfect transcript) is quite low, especially compared with the level of effectiveness of other IR systems using the same collection, whose performance data can be found in the TREC Proceedings, for example. The reasons for this behaviour are due to the fact that in our experiments the IR system's parameters have not been fine-tuned for the collection used and no precision enhancement technique, like for example the use of noun phrases, is employed, as it is done in almost all systems taking part in TREC [11]. This is a deliberate choice. It is easy to foresee that the use of such techniques would make the difference between the use of the perfect and imperfect transcripts much bigger, since they are likely to be highly affected by the errors present in the SR transcripts. The use of these techniques, therefore, would not allow an easy analysis of the causes of the loss of effectiveness.

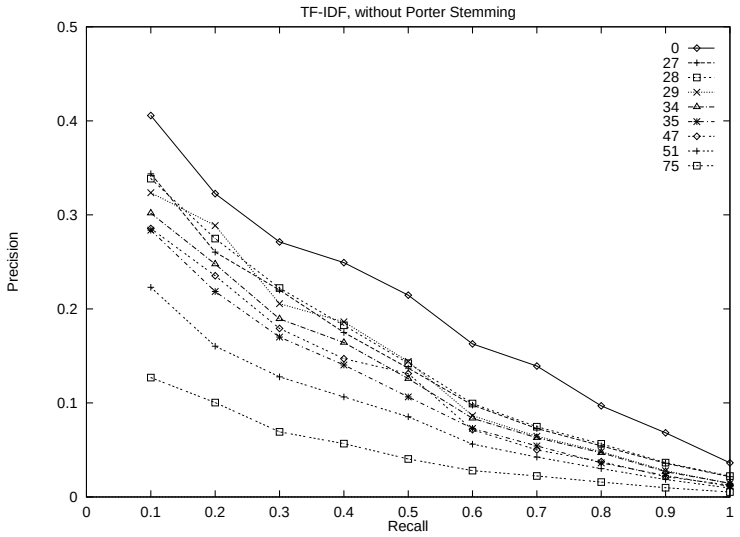

Figure 3. Results using the $tf - idf$ weighting scheme without stemming.

Figure 2 also shows that for WERs ranging from 27% to 47% there is not much difference in effectiveness. Moreover, our IR system seems to perform better with some higher levels of WER than with lower ones: this is not statistically significant. Loss of effectiveness can only be observed at over 50% WER, and significant low levels of effectiveness can be found for 75% WER, where the number of errors in the query is so large that what is left of the original query is not enough for the IR system to work on. We can then conclude that standard IR is quite robust to WRE in spoken queries. One of the possible explanations of this fact can be found in the kind of errors a SR system produces on the query. It is a known fact that SR produces more errors in short words than in long words [21]. Short words are not very useful for IR purposes, since they are mostly non content-bearing words many of which can be found in the stoplist. So, as long as the WER is relatively low, mainly short functional terms are affected. When the WER is higher, longer words are affected too and since these words are generally very important for IR, we have a considerable degradation in the effectiveness of the IR process.

In order to study further the effects of WRE on the effectiveness of SQP, a large number of experiments using the reference IR system were carried out. In these experiments some of the parameters of the IR process were changed to study their effects on the effectiveness on the SQP task in relation to the different levels of WER. Figure 3 shows the effect on IR effectiveness of the removal of the stemming phase of the indexing. Stemming has been proved to generally improve performance in textual IR [19]. Surprisingly, stemming seems to have the opposite effect in SQP, so much that the removal of such a phase actually improves effectiveness. There is no clear explanation for this phenomenon. The effect (either positive or negative) of stemming on the query terms should be very little and should not affect the performance of an IR system, but this is not what these results show.

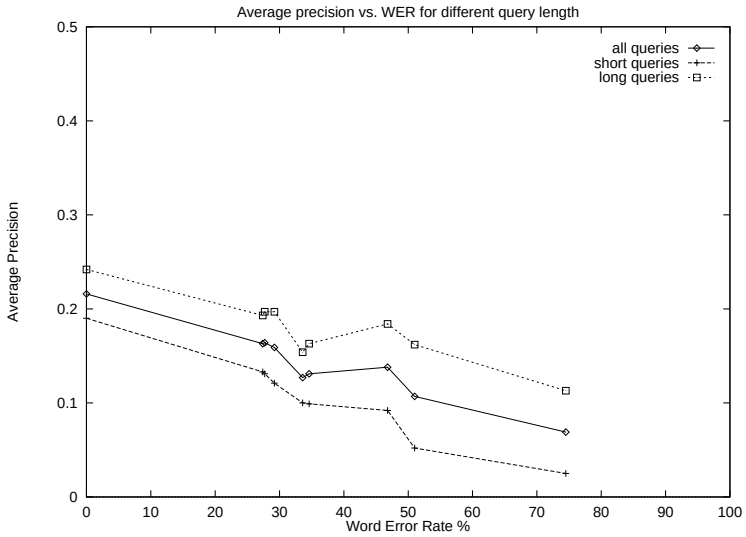

Figure 4. Relation between word recognition errors, average precision and query length.

Another interesting phenomenon was observed when the classic $tf-idf$ weighting scheme was substituted by a weighting scheme that only uses the idf weight. It was surprising to observed that the idf weighting scheme produced the same level of effectiveness than $tf - idf$. This is in contrast to what generally happens in textual IR, where the tf weight is very important [19]. However, since these results can also be observed for the run using the perfect transcript, we could attribute them to idiosyncrasies of the particular collection used and cannot be generalised. Other experiments involving the use of different versions of the tf weighting scheme and of different sizes of stoplists did not produce significantly different results from the ones reported here.

More experimentation, in particular with other collections, is needed to analyse fully the above uncharacteristic phenomena before making any dangerous generalisation.

5.2 Effects of Word Recognition Errors and Query Length

Another series of experiments was conducted to test the robustness of SQP in relation to query length. In fact, it is intuitive to think that the same WER would have a much detrimental effects on short queries than on long ones.

Figures 4 and 5 report average and median precision values for short and long queries at different levels of WER. In this study, short queries are queries with less than 28 terms, and long queries those with more than 28 terms; where 28 terms is the median length of queries. The average number of terms in a query, after stopterm removal is 35, therefore there is a number of considerably

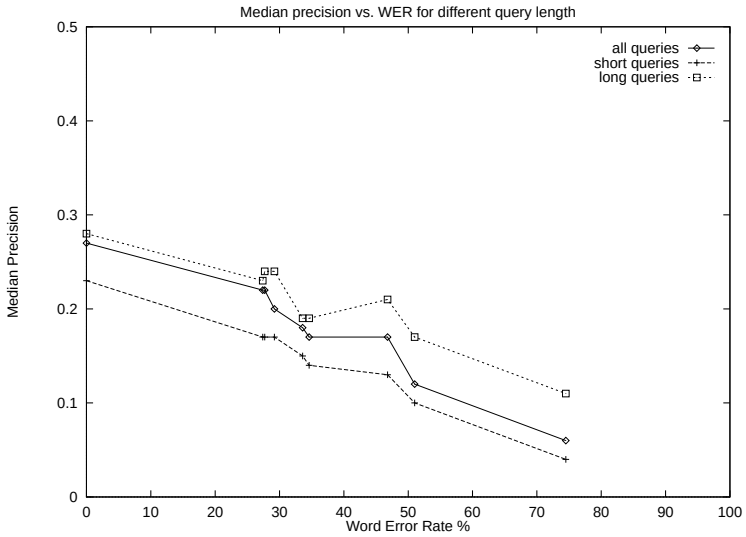

Figure 5. Relation between word recognition errors, median precision and query length.

long queries raising the average. We can notice that short queries have a lower average precision for any level of WER, while long queries have a higher average precisions for any level of WER. This proves the intuition that long queries are more robust to WRE than short queries. However, we should also notice that the median values for all levels of WER are always better than the average values, suggesting that some queries give very bad performance and lower the average. The strange behaviour of the IR system for the 47% WER, that give better performance than some lower WERs, can be explained by the "lucky" correct recognition of one or more important terms than enabled that run to find one or more relevant documents than other runs at lower levels of WER. This event would be ruled out by experiments with larger sets of queries.

6 Effectiveness of Relevance Feedback in Spoken Query Processing

Given the acceptable level of effectiveness of an IR system performing SQP at levels of WER roughly below 40%, we can conclude that it will be quite likely that in the first n retrieved documents (with n dependent on the user's preference and usually less than 10) there will be some relevant ones. It is therefore possible to use both standard and pseudo RF to try to improve the effectiveness of the SR task.

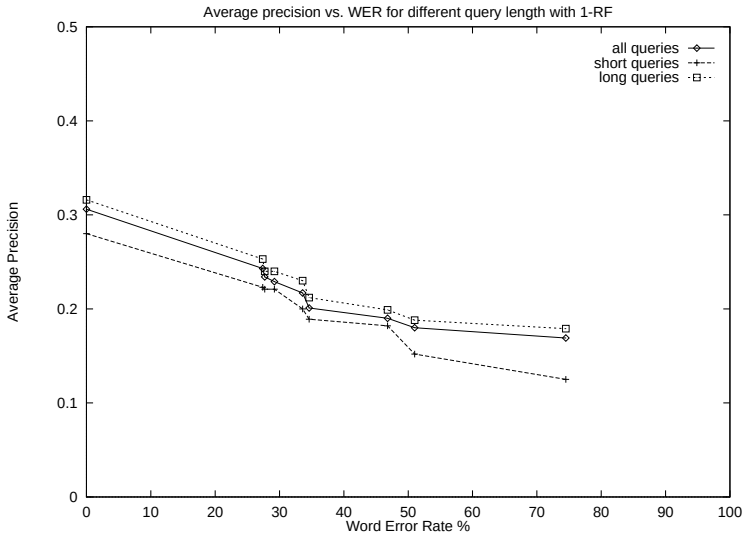

Figure 6. Average precision of standard relevance feedback with 1 relevant documents.

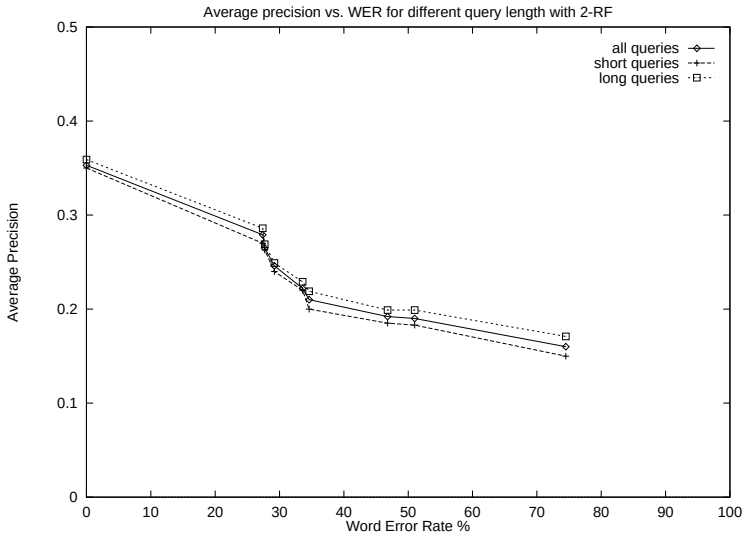

Figure 7. Average precision of standard relevance feedback with 2 relevant documents.

6.1 Standard Relevance Feedback

Standard RF was performed using the first n known relevant documents found in the ranked list of retrieved documents. This process simulates a user manually selecting the relevant documents.

Figures 6 and 7 show the effects of using $n = 1$ and $n = 2$ relevant documents in the standard RF process. Comparing these figures with figure 4 we can observe a significant increase in the effectiveness of the IR system (as measured by average precision) for every level of WER. We can also notice that the dif-

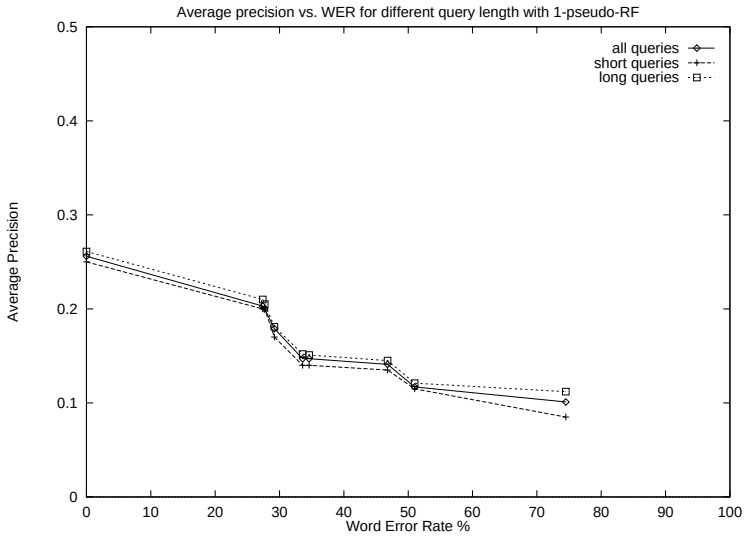

Figure 8. Average precision of pseudo relevance feedback with 1 relevant documents.

ference between effectiveness of short and long queries has almost disappeared; this is particularly true when 2 documents are used in the RF. Both effects are due to the expansion of the original queries with terms extracted from the relevant document(s). The added terms enable the construction of longer queries and the partial recovery of the loss in effectiveness due to WREs in the original query terms.

6.2 Pseudo Relevance Feedback

Pseudo relevance feedback was performed using the first n documents in the ranked list of retrieved documents, irrespective of the actual relevance or not of the documents. This process can therefore be performed without any direct user involvement.

Figures 8 and 9 show the effects of using $n = 1$ and $n = 2$ relevant documents. The same conclusions regarding the increase in effectiveness for any level of WER that were reached for standard RF are also partially valid here. However, one can notice that the effectiveness has not improved as much, and that the difference between short and long queries has effectively disappeared. These effects can be explained by considering that pseudo RF may use documents that are not actually relevant. In fact, an analysis of the data proved that about 30% of the documents used by the pseudo RF where not relevant, on average. This percentage was lower for lower levels of WER and higher for higher levels, due to the lower quality of the ranked list of retrieved documents for higher levels of WER.

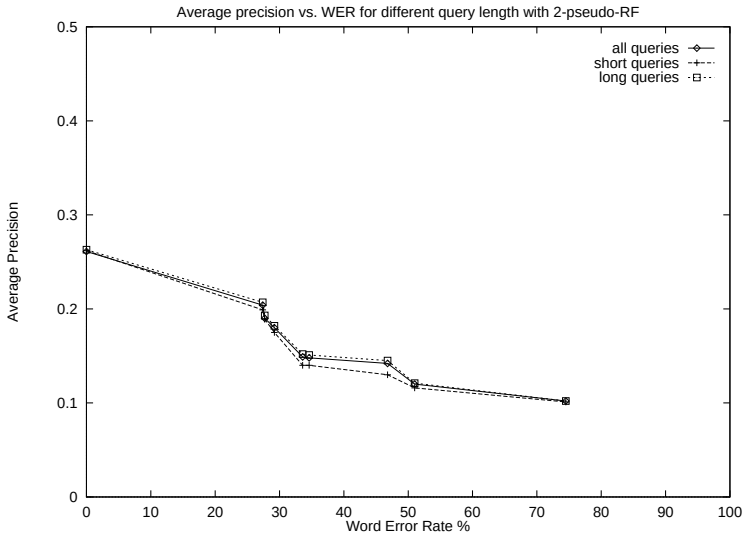

Figure 9. Average precision of pseudo relevance feedback with 2 relevant documents.

7 Limitations

In the previous sections we reported some results of an extensive analysis of the effects of WRE in queries on the effectiveness of an IR system. To the best of our knowledge there is only one other similar study, by Barnett et al. [12]. However, we believe the work presented here to be more complete than Barnett et al., since it uses a larger number of query sets, a larger collection of documents, and a more classical, similarity based, IR system that was not tuned to the test collection used. In addition, the work reported here studies the effectiveness of standard and pseudo relevance feedback on SQP, following an approach similar to that presented by Singhal and Pereira for SDR [7].

Nevertheless, there are at least two important limitations to this study:

(1) The queries used in the experimentation were too long and not really representative of typical user queries. However, it has been long recognised that query length is mainly dependent upon the application domain and the IR environment. Some initial user studies on spoken queries conducted by the author indicates that spoken queries are usually longer than written queries, but further research is needed to support this claim.
(2) The WERs of the queries used in this experimentation were typical of "dictated" spoken queries, since this was the way they were generated. Dictated speech is considerably different from spontaneous speech and easier to recognise [22]. We should expect spontaneous spoken queries to have higher levels of WER and different kinds of errors. Unfortunately, there is no set of spontaneous spoken queries available for SQP experimentation and its construction is not an easy task.

Further empirical work is required to clarify the ways in which spontaneous queries differ in length and nature from dictated ones. Future work will be directed towards overcoming some of these limitations.

8 Other Techniques for Improving Spoken Query Processing

We are currently experimenting with techniques aimed at improving the effectiveness of SQP for IR. The presentation of these techniques is outside the scope of this article, so here we will only outline the directions of this work in progress.

8.1 Use of Prosodic Stress for Topic Detection

In [23] we reported on an investigation into the relationship between acoustic stress in spoken sentences and information content. Using a set of spontaneous sentences from the OGI Corpus, the *average acoustic stress* was measured for each word throughout each utterance. Using the manual transcripts of the same sentences, an IR index (the $tf - idf$ weight) was calculated for each word. In the vast majority of the utterances analysed, the scatter plot of the two measures showed a correlation between high values of average acoustic stress and high values of the IR index of the words. Another proof of such a relationship was derived from the histogram of the words with high average acoustic stress vs. the IR index. It confirmed that a word with high average acoustic stress has also a high value of the IR index and, if we trust IR indexes, it also has high informative content.

This study confirmed our hypothesis of a direct relationship between acoustic stress and information content (as identified by IR weighting) in spontaneous spoken sentences. The next stage of this work will be the integration of prosodic stress and IR weighting evidence into an new IR weighting algorithm for spontaneous speech. This weighting schema will take into account both acoustic and statistical clues to characterise the document/query informative content. It will be extremely useful in a number of tasks and in particular for true spontaneous SQP, where the short length of queries requires every possible clue to fully capture the user's information need.

8.2 Combination of Semantic and Phonetic Term Similarity

As already indicated, a fundamental problem of IR is term mismatch. A query is usually a short and incomplete description of the user's information need.

Term mismatch produces an incorrect relevance ranking of documents with regards to the information need expressed in the query [24]. A similar problem, term misrecognition, can be found in SDR and SQP, where terms misrecognised by the speech recognition process are found not matching in query and document representations. As we have seen for spoken queries, this hinder the effectiveness of the IR system.

In [25] we presented a model for dealing with the term mismatch and the term misrecognition problems in SDR and SQP. Term similarity is used at retrieval time to estimate the relevance of a document in response to a query by looking not only at matching terms, but also at non-matching terms whose semantic and/or *phonetic similarity* are above a predefined threshold. Semantic similarity can help solve the term mismatch problem. It can be estimated using Expected Mutual Information Measure or some similar measure. Phonetic similarity, on the other hand, can help tackle the term misrecognition problem. It can be estimated using Error Recognition Confusion Matrices, for example. An experimental investigation is currently being carried out. The experimental results will provide useful feedback on the effectiveness of the models proposed and on how to effectively combine semantic and phonetic similarity.

9 Conclusions

This paper reports on an experimental study on the effects of WRE on the effectiveness of SQP for IR. Despite the limitations of the experimentation presented here, the results show that the use of classical IR techniques for SQP is quite robust to considerably high levels of WER (up to about 47%), in particular for long queries. Moreover, both standard RF and pseudo RF enable to improve the effectiveness of SQP, in particular for short queries.

Acknowledgements

Most of the work reported in this paper was carried out while the author was a visiting fellow at the International Computer Science Institute, Berkeley, CA, USA.

The author wishes to thank Jim Barnett, Stephen Anderson, and Dragon Systems for generating and providing the spoken queries used in this study.

References

[1] J. Garofolo, C. Auzanne, E. Voorhees, The TREC spoken document retrieval track: a success story, in: Proceedings of the TREC Conference, Gaithersburg, MD, USA, 1999, pp. 107–130.

[2] A. Singhal, J. Choi, D. Hindle, D. Lewis, F. Pereira, AT&T at TREC-7, in: Proceedings of the TREC Conference, Washington DC, USA, 1998, pp. 239–253.

[3] F. Crestani, Vocal access to a newspaper archive: design issues and preliminary investigation, in: Proceedings of ACM Digital Libraries, Berkeley, CA, USA, 1999, pp. 59–68.

[4] A. Tombros, M. Sanderson, Advantages of query biased summaries in Information Retrieval, in: Proceedings of ACM SIGIR, Melbourne, Australia, 1998, pp. 2–10.

[5] A. Tombros, F. Crestani, Users's perception of relevance of spoken documents, Journal of the American Society for Information Science 51 (9) (2000) 929–939.

[6] C. Cleverdon, J. Mills, M. Keen, ASLIB Cranfield Research Project: factors determining the performance of indexing systems, ASLIB (1966).

[7] A. Singhal, F. Pereira, Document expansion for speech retrieval, in: Proceedings of ACM SIGIR, Berkeley, CA, USA, 1999, pp. 34–41.

[8] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[9] J. Xu, Solving the word mismatch problem through automatic text analysis, Ph.D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, MA, USA (May 1997).

[10] C. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, UK, 1979.

[11] E. Voorhees, D. Harman, Overview of the seventh text retrieval conference (TREC-7), in: Proceedings of the TREC Conference, Gaithersburg, MD, USA, 1998, pp. 1–24.

[12] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, S. Kuo, Experiments in spoken queries for document retrieval, in: Eurospeech 97, Vol. 3, Rodhes, Greece, 1997, pp. 1323–1326.

[13] M. Sanderson, Word sense disambiguation and Information Retrieval, PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK (1996).

[14] F. Crestani, M. Sanderson, M. Theophylactou, M. Lalmas, Short queries, natural language, and spoken document retrieval: experiments at Glasgow University, in: Proceedings of the TREC Conference, Washington D.C., USA, 1998, pp. 667–686.

[15] D. Harman, Ranking algorithms, in: W. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: data structures and algorithms, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992, Ch. 14.

[16] D. Harman, Relevance feedback and other query modification techniques, in: W. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: data structures and algorithms, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992, Ch. 11.

[17] S. Robertson, K. Sparck Jones, Simple, proven approaches to text retrieval, Tech. Rep. TR356, Computer Laboratory, University of Cambridge, UK (May 1997).

[18] F. Crestani, Effects of word recognition errors in spoken query processing, in: Proceedings of the IEEE ADL 2000 Conference, Washington DC, USA, 2000, pp. 39–47.

[19] W. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: data structures and algorithms., Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[20] M. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[21] J. Markowitz, Using speech recognition, Prentice Hall, Upper Saddle River, NJ, USA, 1996.

[22] E. Keller (Ed.), Fundamentals of Speech Synthesis and Speech Recognition, John Wiley and Sons, Chichester, UK, 1994.

[23] R. Silipo, F. Crestani, Prosodic stress and topic detection in spoken sentences, in: Proceedings of the SPIRE 2000, the Seventh Symposium on String Processing and Information Retrieval, La Corunna, Spain, 2000, pp. 243–252.

[24] F. Crestani, Exploiting the similarity of non-matching terms at retrieval time, Journal of Information Retrieval 2 (1) (2000) 23–43.

[25] F. Crestani, Combination of semantic and phonetic term similarity for spoken document retrieval and spoken query processing, in: Proceedings of the 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Madrid, Spain, 2000, pp. 960–967.