Effects of Word Recognition Errors in Spoken Query Processing

Fabio Crestani*
International Computer Science Institute
1947 Center St. Suite 600
Berkeley, CA 94704, USA
fabioc@icsi.berkeley.edu

Abstract

The effects of word recognition errors (WRE) in Spoken Document Retrieval have been well studied and well reported in recent Information Retrieval (IR) literature. Much less experimental work has been devoted to studying the effects of WRE in Spoken Query Processing in IR. It is easy to hypothesize that given the typical length of the user query, the effects of WRE in spoken queries on the performance of IR systems must be destructive. The experimental work reported in this paper intends to test that. The paper reports on the background of such a study, on the construction of a suitable test collection, on the first experimental results obtained and on the limitations of the study. The results show that classical IR techniques are quite robust to considerably high levels of WRE rates in spoken queries (roughly below 40%), in particular for long queries.

1 Introduction

The effects of word recognition errors (WRE) in spoken documents on the performance of an Information Retrieval (IR) system 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 [9]). In the context of the SDR track participants carry out a number of retrieval runs on a collection of spoken documents. The collection is relatively large from a speech recognition perspective, but small from an IR perspective. Almost invariably these documents are processed by a speech recognition (SR) system and transcripts for these documents are fed to a classical (textual) IR system. Naturally, given the limitations of current automatic SR technology, these transcripts will contain a number of errors that will make the transcripts differ from the perfect (human generated) transcript. Participants of the SDR track are also provided with a number of textual queries, and a typical retrieval run consists in retrieving a number of spoken documents in response to a textual query using the document transcripts. The indexes that the IR system uses are generated from the SR transcripts. In many aspects a SDR run is similar to an "ad hoc" run (i.e. a standard retrieval of textual documents in response to textual query), since documents have to be ranked by their evaluated relevance to a query, using a representation of the document and query content that is dependent upon the model used by the IR system. The main difference between an ad hoc run and a SDR runs is in the quality of the document representation, and therefore of the indexes, used by the IR system. While in an ad hoc run document representations are "certain", in a SDR run they are "uncertain", and they may differ considerably from the reality (i.e. the perfect transcript) depending on the quality of the SR process.

This additional uncertainty in the IR process has been tackled in many different ways by TREC SDR participants [15, 14, 6, 1]. The most effective techniques employed make use of various forms of document expansion (see for example [15, 6]). However, it has been noted 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 characterization of the document content are often repeated several times, the probability that a term will always be misrecognized is quite low and there is a good chance that a term will be correctly recognized at least once. Variations of classical IR weighting schemes (for example giving lesser importance to the within document term frequency) that are able to cope with reasonable levels of WER have been proposed [15], but these solutions were found not effective for short documents.

Very little research work has been devoted to studying the *effects of WRE in Spoken Query Processing* (SQP). A spoken query can be considered similar to a very short doc-

*Current author's address: Department of Computer Science, University of Strathclyde, Glasgow G1 1XH, Scotland, UK. Email: fabioc@cs.strath.ac.uk

ument and high levels of WER may have devastating effects on the performance of the IR system. In a query, like 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. Techniques making use of automatic query expansion based on semantic term similarity [20] may not be useful, given the uncertainty associated to the terms present in the query representation and upon which the query expansion should be based. The combination of semantic and phonetic similarities for query expansion is currently being investigated [5].

In this paper we present the results of an experimental study of the effects of WRE in spoken queries on the effectiveness of a SQP system. To the best of my knowledge there is only one other similar study, by Barnett et al. [2]. This work is more complete than Barnett et al. and uses a more classical, similarity-based, IR system that was not fine-tuned to the test collection used. I believe the results of this study are more general and generalizable than those reported by Barnett et al..

The paper is structured as follows. Section 2 describes the background and the motivations of the work. In this context, section 3 points out some important research issues. Some of these issues will be addressed in the remainder of the paper by means of an experimental analysis. Section 4 presents the experimental environment of the analysis, the results of which are reported and discussed in section 5. Section 6 proposes some techniques for overcoming some of the problems found with dealing with spoken queries and high levels of WER. The conclusions of the study and directions of future work are reported in section 7.

2 Information Retrieval and Speech: 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 telephone line. A first prototype system, called *Interactive Vocal Information Retrieval System* (IVIRS), is currently being developed [4]. 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.

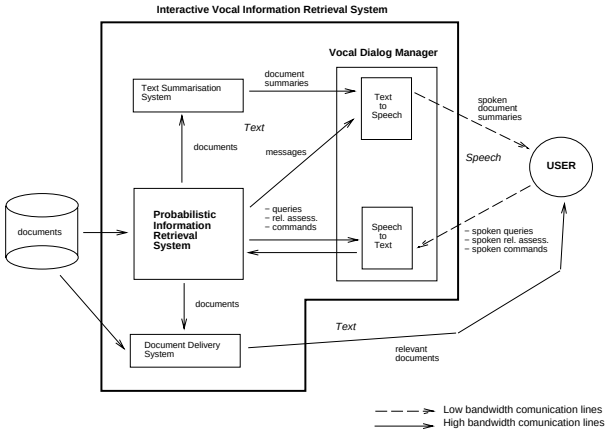

Figure 1. Schematic view of the IVIRS prototype

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 sent to him via the Document Delivery System.

A prototype implementation of IVIRS enabling journalist to access newspapers archives while out on an assignment, is currently in progress [4]. A “divide and conquer” approach has been followed, consisting of dividing the implementation and experimentation of IVIRS in the parallel implementation and experimentation of different components. Work carried out so far has concentrated on implementing and experimenting with the Document Summarization System [19], the Text-to-Speech and Speech-to-Text modules of the VDM [18], and the Document Delivery Sys-

tem.

3 Issues Related to the Use of Spoken Queries in Information Retrieval

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. As already mentioned, a number of studies have been devoted to studying the effects of WRE in SDR, but 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 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 SR system is operating. It is not unusual to require 200 time units to process one time unit of speech [9]. This does not constitute a serious problem in SDR, where spoken document are processed off-line to produce transcripts and the 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 [3]. Therefore it is advisable that this time is kept short, in the order of 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 [16]. 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 [7]. 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 [21]. 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). 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 The Experimental Environment

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

1. a collection of textual documents;

Data sets:	WSJ 1990-92
Num. of doc.	74.520
Size in Mb	247
Num. of queries	35
Unique terms in doc.	123.852
Unique terms in queries	3.304
Avg. doc. length	550
Avg. doc. length (unique terms)	180
Avg. query length (with stopterms)	58
Avg. query length (without stopterms)	35
Med. query length (without stopterms)	28
Avg. num. of rel. doc. per query	30

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

2. a set of spoken queries recognized at different levels of WER with associated relevance assessments;

3. an IR system.

The next sections report on the characteristics of the data and system used in this experimentation.

4.1 The Document Collection

Since to the best of my knowledge there is no test collection available with spoken queries, it was necessary to generate it from an existing textual collection. The document collection we used is the *TREC-5 B*, a subset of the collection generated for TREC 5 [10]. 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.

In this work 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; the text of all sections of the document was considered indistinctively.

4.2 Spoken Queries

We used a set of 35 queries (topics 101-135 of TREC 5) with the corresponding set of relevant documents. Some of the fields of the query were not used in the experiments reported in this paper. In fact, the only fields used were title, description, and concepts. The text in these fields was considered indistinctively. This made the queries short enough to be a somewhat more realistic examples of "real" user queries.

Since the original queries were in textual form, it was necessary to produce then in spoken form and have then recognized by a SR system. This work was carried out by Jim Barnett and Stephen Anderson of Dragon Systems Inc.

Barnett and Anderson had one single (male) speaker dictate the queries. The spoken queries were then recognized by Dragon's research LVCSR system, a SR system that has a 20.000 vocabulary and a bigram language model trained on the Wall Street Journal language model. By altering the width of the beam search[1], transcripts at different levels of WER were generated. More details on the actual process used in generating these different sets of transcripts are not important for the experimentation reported in this paper and can be found in [2]. Different sets of transcripts for the query set were generated. The error characteristics of these sets of transcripts (we shall refer to them as query sets) are reported in table 2.

Notice that each query set has been denoted with a name that refers to the approximated average WER of that set (i.e. the "Avg. % Error"). The set identified by 0 (not reported in the table, but reported in the figures as a reference line) is the perfect transcript.

4.3 The Information Retrieval System

The system used for the work reported in this paper is 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 multitasking is eminently suitable for supporting such a modular architecture. This system was used by the Glasgow IR group for submissions to several TREC conferences, 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:

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

[1] The beam width was chosen as the major parameter to alter because it was believed that this yields relatively realistic recognition errors.

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

Table 2. Characteristics of the different query sets.

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, such as the dot product, to the document and query representation 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 classic formula depending on the way the tf and idf weights are computed [8]. We chose this one because it is the weighting scheme we are most familiar with. Other weighting schemes may prove to be more or less effective.

4.4 The Evaluation Methodology

The main IR effectiveness measures are Recall and Precision. *Recall* (R) is defined as the portion of all the relevant documents in the collection that has been retrieved. *Precision* (P) is the portion of retrieved documents that is relevant to the query. These values are often 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 remind that, experimentally, these measures have proved to be related in such a way that high precision brings low recall and vice versa. In other words, if one desires high precision, has to accept low recall, and vice versa.

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 and precision and recall values were obtained. The results reported in the following tables and graphs are averaged over the entire sets of 35 queries.

5 Effects of Word Recognition Errors in Spoken Queries on the Effectiveness of a SQP System

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, only the most interesting ones.

5.1 Effects of WRE on a Standard IR System Configuration

The first analysis was directed towards studying the effects of different WERs in spoken queries on the effectiveness of an IR system using a standard text-based parameters configuration. The parameters configuration most commonly used in 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. Terms appearing in the stoplist are removed. 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. In the experiments reported here a standard stoplist [8] and the stemming and conflation algorithm commonly known as "Porter algorithm" were used [12]. Figure 2 depicts the effects of different WERs in queries on the effectiveness of the IR system using the above standard text-based configuration.

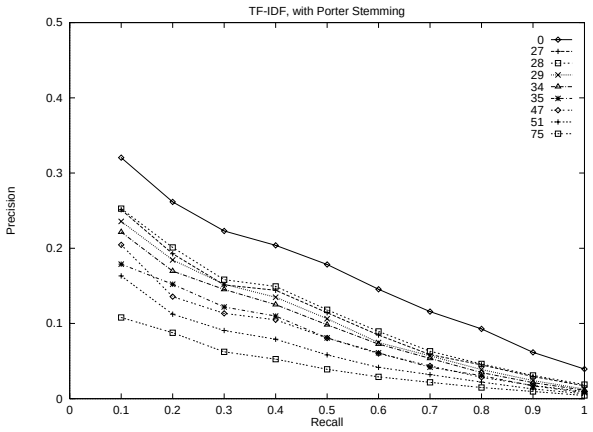

Figure 2. Results using the $\mathrm{tf} - \mathrm{idf}$ weighting scheme and the Porter stemming.

Naturally, it can be noted that the best results are ob-

tained for the perfect transcript (the transcript 0), and there is a degrade in effectiveness that is related to the WER. 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. The reason for these results is due to the fact that no precision enhancement technique, like for example the use of noun phrases or pseudo relevance feedback, was employed in the experimentation reported in this paper. This is a deliberate choice, since it was felt that the use of such techniques would not have allowed a "clean" analysis of the effects of WRE on IR effectiveness.

Figure 2 also shows that for WERs ranging from 27% to 47% there is not much difference in effectiveness. The little difference that can be seen in the figure is not statistically significant. Moreover, some higher levels of WER seem to do better that lower ones; this again is not statistically significant. Serious losses of effectiveness can only be observed at over 50% WER. We can then conclude that the standard IR is quite robust to WRE in spoken queries.

5.2 Effects of WRE on Different IR System Configurations

In order to study 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 [8]. 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. A deeper analysis needs to be carried out to study this effect, looking at single query terms, before these conclusions could be generalized. This will be the object of future work.

Another interesting phenomenon can be observed in figures 4 and 5. Here the classic $tf - idf$ weighting scheme was substituted by a weighting scheme that only uses collection-wide information, the idf weight, and where the frequency of occurrence of a term in a document is not considered. It is again surprising to observe that the idf weighting scheme produces the same level of effectiveness

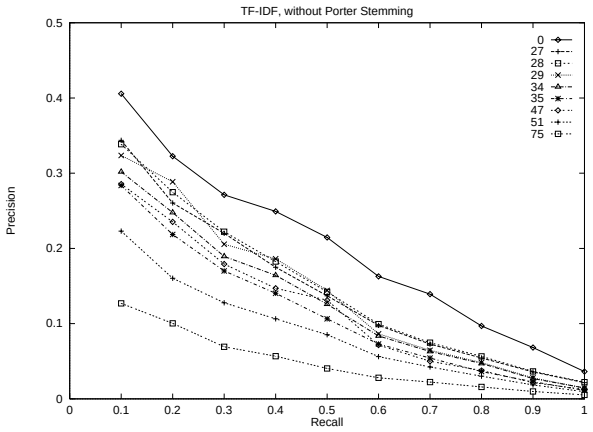

Figure 3. Results using the $\mathrm{tf} - \mathrm{idf}$ weighting scheme and no stemming.

than $tf - idf$. This is in contrast to what generally happens in textual IR, where within document frequency constitutes important information for the weighting scheme [8]. Moreover, figure 5 confirms that the use of stemming is detrimental to the effectiveness of an IR system in SQP, as already observed previously in figure 3. 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. More experimentation, in particular with other collections, is needed to analyze fully this phenomenon before making any dangerous generalization.

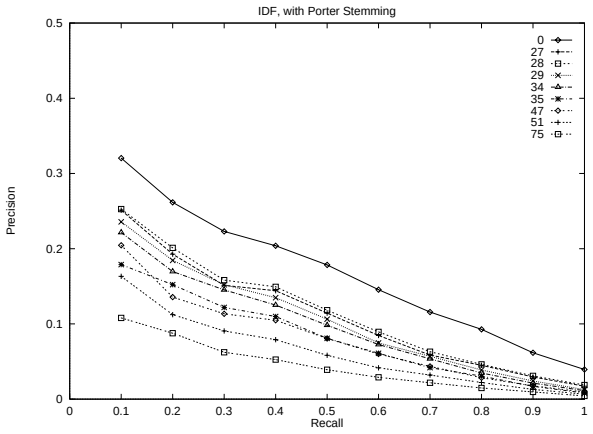

Figure 4. Results using the idf weighting scheme and the Porter stemming algorithm.

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 and will not be presented.

Query sets	0	27	28	29	34	35	47	51	75
10	0.40	0.34	0.33	0.32	0.30	0.28	0.28	0.22	0.12
20	0.32	0.26	0.27	0.28	0.24	0.21	0.23	0.16	0.10
30	0.27	0.22	0.22	0.20	0.18	0.17	0.17	0.12	0.06
40	0.24	0.17	0.18	0.18	0.16	0.14	0.14	0.10	0.05
50	0.21	0.13	0.14	0.14	0.12	0.10	0.13	0.08	0.04
60	0.16	0.09	0.09	0.08	0.08	0.07	0.07	0.05	0.02
70	0.13	0.07	0.07	0.06	0.06	0.05	0.05	0.04	0.02
80	0.09	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.01
90	0.06	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.00
100	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00
Avg. Prec.	0.22	0.16	0.16	0.16	0.13	0.13	0.14	0.11	0.07

Table 3. Precision values for standard levels of recall for the different query sets, using the "idf-no-porter" weighting scheme.

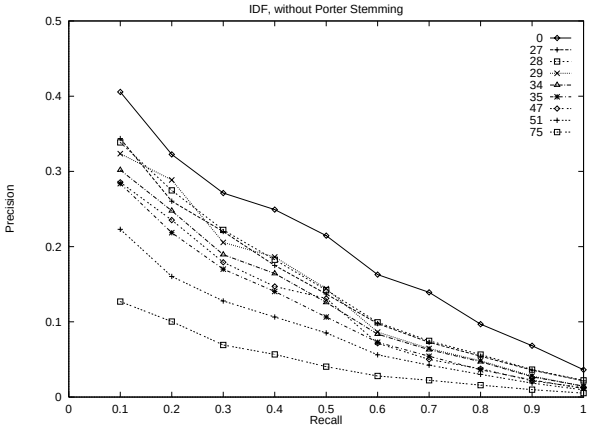

Figure 5. Results using the idf weighting scheme and no stemming.

Table 3 reports the best results obtained in the experimentation, which are obtained with the idf weighting scheme without the use of stemming and with a standard stoplist. We can notice that for levels of WERs ranging from 27% to 47% there is no significant difference in performance. 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 to work on.

One of the possible explanations of the fact that classical IR techniques are considerably robust to high levels of WER, can be found in the kind of errors that a SR system produces on the query. It is a known fact that SR produces more errors in short words than in long words. 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, mostly short functional terms are affected. When the WER is higher, longer words are affected too and since these words are generally very important in IR, we have a considerable degradation in the effectiveness of the IR process.

5.3 Effects of WRE and Query Length

Another series of experiments was conducted to test the robustness of the IR process in relation to query length. It is intuitive to think that the same WER would have a much more detrimental effect on short queries than on long ones. A 50% WER means that on average half of the terms in a query are misrecognized. So, if a query is 20 terms long, only about 10 terms are correct. These 10 remaining terms could still be sufficient to identify relevant documents, as long as the misrecognized terms do not change the query too much. We can imagine the effect of a 50% WER to be higher if the query is only 10 terms long. There might be not enough information in the 5 correctly recognized terms to identify relevant documents.

Table 4 reports the average precision values for short and long queries and compare these data with the overall average precision at different levels of WER. 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 a query. The average number of terms in a query, after stopterm removal is 35, therefore there are a number of considerably 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. The strange behavior of the IR system for the 47% WER, that gives better performance than some lower WERs, can

Query sets	0	27	28	29	34	35	47	51	75
Avg. Prec. overall	0.22	0.16	0.16	0.16	0.13	0.13	0.14	0.11	0.07
Avg. Prec. short q.	0.19	0.13	0.13	0.12	0.10	0.10	0.09	0.05	0.03
Avg. Prec. long q.	0.24	0.19	0.20	0.20	0.15	0.16	0.18	0.16	0.11

Table 4. Average precision values for the different query sets, using long or short queries with the "idf-no-porter" weighting scheme.

be explained by the correct recognition of one or more important terms that enabled the IR system to find one or more relevant documents than with queries at lower levels of WER. This event should be considered not uncommon and can only be ruled out by experiments with larger sets of queries.

Table 5 reports also the median precision values for short and long queries and compare these data with the overall median precision at different levels of WER. We can observe here a similar better behavior of the IR system with long and short queries. However, we should 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 as to lower the average. It will be necessary to exploit other techniques to improve the performance of the IR system for these queries. The next section provide some indications of what techniques could be used to improve the performance of a SQP system.

6 Techniques for Improving SQP

Given the acceptable level of effectiveness of 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. In a previous study we have tested the ability of the user to understand if a document is relevant to the user's information need when the document is presented in the form of a short spoken summary [18]. This scenario is consistent with the goals of the SIRE project. That study showed that the user is indeed able to perceive the relevance of a document presented in the form of a spoken summary. This result, together with the results reported in this paper enable us to conclude that *relevance feedback* [8] could be a good strategy to improve effectiveness in a SQP task. The user could find at least one relevant document and feed it back to the IR system which will expand the initial query (therefore also recovering some of the problems due to short queries) and enable to find more relevant documents. Some technique can also be used to recover some of the WRE in the original query by means of the information provided by the relevance feedback.

Another finding of the study reported in this paper is that long queries are much more robust to high levels of WER than short queries. For this reason, in the design of the VDM for the IVIRS we will have to exploit dialogue techniques that will elicit the longest possible queries from the users. This is consistent with results of other projects (see for example [11]), and there exists already a number of techniques that we might be able to use in this context [17].

7 Conclusions and Future Work

This paper reports on an experimental study on the effects of WRE on the effectiveness of SQP system. The results show that classical IR techniques are quite robust to considerably high levels of WER (up to about 40%), in particular for long queries.

However, the experimentation reported here falls short in a number of ways:

- The queries used are too long and not really representative of typical user queries (although some initial unpublished user studies on spoken queries indicates that spoken queries are usually longer than written queries).
- 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 recognize. We should expect spontaneous spoken queries to have higher levels of WER and different kinds of errors.
- The queries used in this experimentation were generated artificially from queries spoken in a laboratory environment. It is known that telephone speech is more difficult to recognize than laboratory speech. In addition, transcripts from telephone speech have different types of errors than laboratory speech. This experimentation needs to be repeated with queries that are closer to the operative conditions of IVIRS.

Future work will be directed towards overcoming some of the above limitations. Moreover, future work will investigate the use of relevance feedback as a way of recovering

Query sets	0	27	28	29	34	35	47	51	75
Med. Prec. overall	0.27	0.22	0.22	0.20	0.18	0.17	0.17	0.12	0.06
Med. Prec. short q.	0.23	0.17	0.17	0.17	0.15	0.14	0.13	0.10	0.04
Med. Prec. long q.	0.28	0.23	0.24	0.24	0.19	0.19	0.21	0.17	0.11

Table 5. Median precision values for the different query sets, using long or short queries with the "idf-no-porter" weighting scheme.

WRE and as a way of improving the effectiveness of the IR process with spoken queries. The use of relevance feedback in a sonified IR environment will also be studied in conjunction with spoken dialogue techniques for eliciting longer queries from the user. We expect that the combination of relevance feedback and longer queries will produce considerably better results for any level of WER.

Acknowledgments

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] D. Abberley, S. Renals, G. Cook, and T. Robinson. The THISL spoken document retrieval system. In *Proceedings of the TREC Conference*, Gaithersburg, MD, USA, Nov. 1997.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. Kuo. Experiments in spoken queries for document retrieval. In *Eurospeech 97*, volume 3, pages 1323–1326, Rodhes, Greece, Sept. 1997.

[3] C. Cleverdon, J. Mills, and M. Keen. *ASLIB Cranfield Research Project: factors determining the performance of indexing systems*. ASLIB, 1966.

[4] F. Crestani. Vocal access to a newspaper archive: design issues and preliminary investigation. In *Proceedings of ACM Digital Libraries*, pages 59–68, Berkeley, CA, USA, Aug. 1999.

[5] 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, July 2000. In press.

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

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

[8] W. Frakes and R. Baeza-Yates, editors. *Information Retrieval: data structures and algorithms*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[9] J. Garofolo, E. Voorhees, C. Auzanne, V. Stanford, and B. Lund. 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of the TREC Conference*, pages 79–90, Gaithersburg, MD, USA, Nov. 1998.

[10] D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *Proceedings of the TREC Conference*, Gaithersburg, MD, USA, Nov. 1996.

[11] J. Peckham. Speech understanding and dialogue over the telephone: an overview of the ESPRIT SUNDIAL project. In *Proceedings of the Workshop on Speech and Natural Language*, pages 14–27, Pacific Groce, CA, USA, Feb. 1991.

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

[13] M. Sanderson. *Word Sense Disambiguation and Information Retrieval*. PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, 1996.

[14] M. Siegler, M. Witbrock, S. Slattery, K. Seymore, R. Jones, and A. Hauptmann. Experiments in spoken document retrieval at CMU. In *Proceedings of the TREC Conference*, Gaithersburg, MD, USA, Nov. 1997.

[15] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira. AT&T at TREC-7. In *Proceedings of the TREC Conference*, pages 239–253, Washington DC, USA, Nov. 1998.

[16] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of ACM SIGIR*, pages 34–41, Berkeley, CA, USA, Aug. 1999.

[17] R. Smith and D. Hipp. *Spoken natural language dialog systems: a practical approach*. Oxford University Press, Oxford, UK, 1994.

[18] A. Tombros and F. Crestani. Users's perception of relevance of spoken documents. Technical Report TR-99-013, International Computer Science Institute, Berkeley, CA, USA, July 1999.

[19] A. Tombros and M. Sanderson. Advantages of query biased summaries in Information Retrieval. In *Proceedings of ACM SIGIR*, pages 2–10, Melbourne, Australia, Aug. 1998.

[20] E. Voorhees. On expanding query vectors with lexically related words. In *Proceedings of the TREC Conference*, pages 223–232, Gaithersburg, MD, USA, Nov. 1993.

[21] 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.