INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

An Experimental Study of the Effects of Word Recognition Errors in Spoken Queries on the Effectiveness of an Information Retrieval System

Fabio Crestani[*]

TR-99-016

October 1999

Abstract

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 reported in recent IR literature. Most of the research in this direction has been promoted by the Spoken Document Retrieval track of TREC. Much less experimental work has been devoted to studying the effects of WRE in spoken queries. It is easy to imagine that given the typical length of the user query, the effects of WRE in queries on the performance of an IR system 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 test collection, and on the first experimental results. The preliminary conclusions drawn from the experimentation enable to give some useful indications for the design of spoken query systems, despite the recognized limitations of the study.

[*] International Computer Science Institute, 1947 Center St. Suite 600, Berkeley, CA 94704, USA, email: fabioc@icsi.berkeley.edu.

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 [31, 10].

In the context of the SDR track participants are asked to 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 system and transcripts for these documents are fed to a classical (textual) IR system. Naturally, given the limitations of current automatic speech recognition technology, these transcripts will contain a number of WRE that will make this transcripts differ from the perfect (human generated) transcript. Participants to the SDR track are also provided with a number of textual queries. Therefore, a typical retrieval run consists in retrieving a number of spoken documents in response to a textual query using the automatically generated document transcripts. The indexes that the IR system uses to estimate the Retrieval Status Value (RSV) of documents are generated from the automatically produced transcripts. In many aspects a SDR run is similar to an "ad hoc" run, 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 IR model used by 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 speech recognition process.

The additional uncertainty in the IR process due to the uncertain transcription of spoken documents has been tackled in many different ways by TREC SDR participants [24, 25, 23, 6, 1]. The most effective techniques employed make use of various forms of document expansion (see for example [24, 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 time, the probability that a term will always be missrecognized is quite low. 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 [24], but these solution were found not valid for short documents.

Very little research work has been devoted to studying the *effects of WRE in spoken queries*. 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 missrecognition of a term may cause the term to disappear completely from the query representation. Because of this a large set of potentially relevant documents indexed using that term will not be retrieved. Techniques making use of automatic query expansion based on term similarity [32, 15] 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.

In this paper I present the results of an experimental study of the effects of WRE in spoken queries on the effectiveness of a IR system. To the best of my knowledge there is only one other similar study, by Barnett et al. [2]. However, this work is more complete than Barnett et al. and uses a more classical, similarity based, IR system not 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 and 3 describe the background of the work. Section 2 presents different ways in which speech can be combined with IR. Section 3 provides the motivation of this work, explaining the goals of the SIRE project. In this context, section 4 points out some of the issues related to the use of spoken queries in IR. Some of these issues will be addressed in the remainder of the paper by means of an experimental analysis. Section 5 presents the experimental environment of the analysis, the results of which are reported and discussed in section 6. Section 7 proposes some techniques for overcoming some of the problems found with dealing with spoken queries with high levels of WER. The conclusions of the study and directions of future work are reported in section 8.

2 Information Retrieval and Speech

Information Retrieval is the branch of computing science that aims at storing and allowing fast access to a large amount of multimedia information, like for example text, images, speech, and so on [30]. An *Information Retrieval System* is a computing tool that enables a user to access information by its semantic content using advanced statistical, probabilistic, and/or linguistic techniques.

Most current IR systems enable fast and effective retrieval of textual information or documents, in collections of very large size, sometimes containing millions of documents. The retrieval of multimedia information, on the other hand, is still an open problem. Very few IR systems capable of retrieving multimedia information by its semantic content have been developed. Often multimedia information is retrieved by means of an attached textual description.

The marriage between IR and speech is a very recent event. IR has been concerned for the last 50 years with textual documents and queries. It is only recently that multimedia IR has become possible. Progress in speech recognition and synthesis [14] and the availability of cheap storage and processing power have made possible what only a few years ago was unthinkable.

The association between IR and speech has different possibilities:

1. textual queries and spoken documents;

2. spoken queries and textual documents;

3. spoken queries and spoken documents.

The retrieval of spoken documents using a textual query, also known as spoken document retrieval (SDR) is a fast emerging area of research. It involves an efficient and effective combination of the most advanced techniques used in speech recognition and IR. The increasing interest in this area of research is confirmed by the inclusion, for the first time, of a retrieval of spoken documents retrieval track in the TREC-6 conference [31]. The problem here is to devise IR models that can cope with the large number of errors inevitably found in the transcripts of the spoken documents. Research work carried out for retrieval of OCRed documents have proved useful in this context [17], although there are a number of fundamental differences between the two problems. An important aspect of SDR is related to the fact that, although IR models can easily cope with the fast searching of large document collections, fast speech recognition of a large number of long spoken documents is a much more difficult task [16]. Because of this, spoken documents are converted into textual transcripts off-line, and only the transcripts are dealt by the IR system at retrieval time.

Research on the retrieval of textual documents using a spoken query, also known as spoken query retrieval (SQR) is a much more recent and less studied area. Although SQR may seem easier than SDR from a speech recognition perspective, given the smaller size of the speech recognition task involved, from an IR point of view SQR is a more difficult task. While the incorrect or uncertain recognition of an instance of a word in a long spoken document can be compensated by its correct recognition in some other instances in the same document, the incorrect recognition of a word in a spoken query can have disastrous consequences. Queries are generally very short[1] and the failing of recognizing a query word, or worse, the incorrect recognition of a query word will fail to retrieve a large number of relevant documents or wrongly retrieve a large number of non-relevant documents.

The retrieval of spoken documents in response to spoken queries, effectively a combination of SDR and SQR, is a very complex task and is more in the realm of speech processing than IR. In fact, speech recognition and processing techniques have been used to compare spoken words and sentences in their raw form, without the need of generating textual transcripts.

In this paper I will address in an experimental fashion some of SQR problems. The motivation for such a choice are explained in the next section.

[1] There is an on-going debate about realistic query lengths. While TREC queries are on average about 40 words long, Web queries are only 2 words long on average. This recently motivated the creation in TREC of a "short query" track, to experiment with queries of more realistic length.

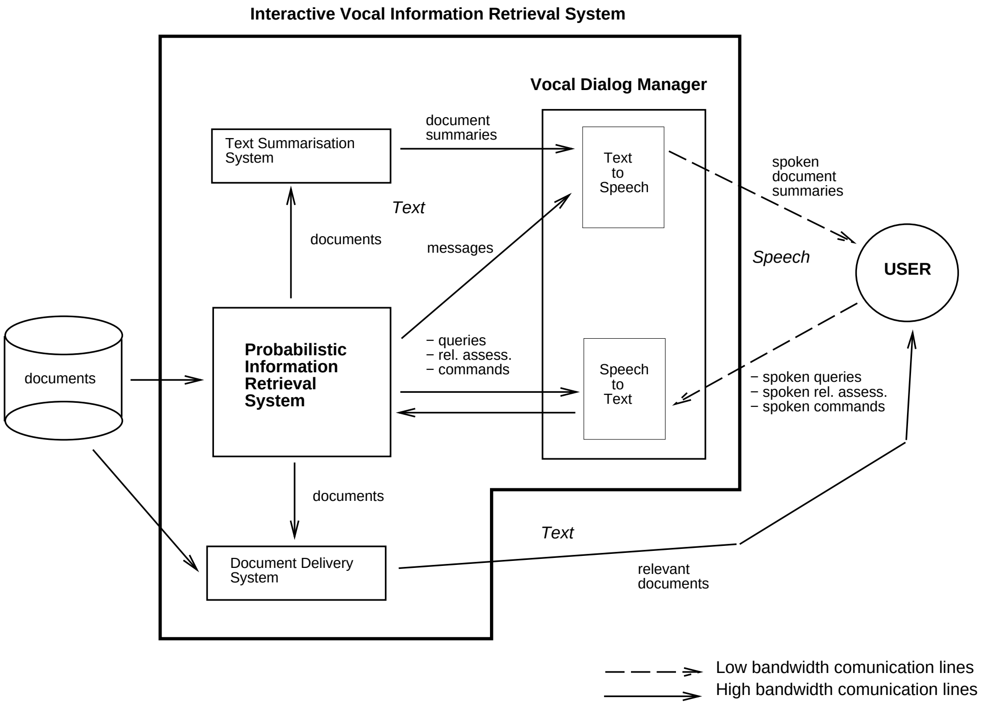

Figure 1: Schematic view of the IVIRS prototype

3 Background: The SIRE Project

The background of the work reported in this paper is related to a project started in the second half of 1997, at Glasgow University and that I am continuing at the International Computer Science Institute. The project is called *SIRE*, which is an acronym for *Sonification of an Information Retrieval Environment*. The project has been partially financed by the Training and Mobility of Researchers (TMR) scheme of the European Commission Fourth Framework of projects and by research fellowship of the International Computer Science Institute.

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. An outline of the system specification of the prototype being built is reported in figure 1.

The prototype system, called *interactive vocal information retrieval system* (IVIRS), is made up of the following components:

- a *vocal dialog manager* (VDM) that provides an "intelligent" speech interface between user and IR system;
- a *probabilistic IR system* (PIRS) that deals with the probabilistic ranking and retrieval of documents in a large textual information repository;

- a *document summarization system* (DSS) that produces a summary of the content of retrieved documents in such a way that the user will be able to assess their relevance to his information need;

- a *document delivery system* (DDS) that delivers documents on request by the user via electronic mail, ftp, fax, or postal service.

It is important to emphasize that such a system cannot be developed simply with off the shelf components. In fact, although some components (DSS, DDS, and the Text-to-Speech module of the VDM) have already been developed in other application contexts, it is necessary to modify and integrate them for the IR task.

The IVIRS prototype 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 (to avoid problems in this phase we devised a login procedure based on keying in an identification number using a touch tone), the user submit a spoken query to the system. The VDM interact with the user to identify the exact part of spoken dialogue that constitutes the query. The query is then translated into text and fed to the PIRS. Additional information regarding the confidence of the speech recognizers is also fed to the PIRS. This information is necessary in order to limit the effects of wrongly recognized words in the query, Additionally, an effective interaction between the system and the user can also help to solve this problem. The system could ask the user for confirmation in case of a uncertain recognition of a word, asking to re-utter the word or to select one of the possible recognized alternatives. The PIRS searches the textual archive and produces a ranked list of documents, and a threshold can be used to find the a set of document 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 DSS that produces a short representation of each document that is read to the user over the telephone by the Text-to-Speech module 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 all together. 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 relevance feedback process is also useful in case of wrongly recognized query words, since the confidence values of query words increase if they are found in relevant documents. This interactive 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 DDS.

The implementation of the prototype system outlined above requires, as a first step, a careful choice of some already existing software components: a speech recognition system, a speech synthesis system, a probabilistic IR system, and a document summarization system. This called for a study of the state-of-the-art of several different areas of research some of which are familiar to us, while others are new to us. Some components were found not to be fully suitable to the task and had to be developed. This is the case of the probabilistic IR system, and a document summarization system. A second step involves the integration of the various components and the development of a model for the VDM and of its interaction with the other components. Finally, the prototype implementation of the overall system requires a careful tuning and testing with different users and in several different conditions (noisy environment, foreign speaker, etc.).

The prototype implementation of IVIRS is still 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. I have implemented and experimented with the DSS [29], the Text-to-Speech and Speech-to-Text modules of the VDM [28], and the DDS and I am currently developing a PIRS able to cope with the different forms of uncertainty involved in the retrieval task.

4 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 spoken documents, but much less research has addressed the effects of WRE in spoken queries. It has to be recognized that SQR 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.

I shall now briefly explain these additional challenges.

In SDR, spoken document are almost always processed off-line using speech recognition techniques . This is due to the computationally intensive nature of the speech recognition process. The time required by a speech recognition system to process a spoken document depends on the system and on the machine the system is operating. It is not unusual to have system requiring 200 time units to process one time unit of speech [10]. 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 document indexes. SQR, 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 is dependent also upon the time the user spends in 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 oder 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 system performing SQR.

The second issue is related to the effectiveness of SQR. It is a well known fact in textual IR that short queries are less effective than long queries in finding relevant documents [26]. This is in large part due to the so called "term-mismatch problem". The cause 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 because, as queries get longer, there is a higher chance of some important term co-occurring in the query and the relevant documents [33]. 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, as the following example shows. The term-mismatch problem becomes more severe when it is combined with the "term-missrecognition" problem. As we already observed, the missrecognition of a spoken query term may cause the term to disappear completely or be replaced by a different term in the query representation. Because of the term missrecognition 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 system 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 parameters of the speech recognition process can be tuned on the data. This testing and tuning of the system is almost always done in SDR. However, this may not be possible with SQR, since it may be the first time the user submit a query (so no previous data 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 environment or using a different microphone). The lack of test data may cause the performance of the speech recognition process to be poor and spoken queries may have a high WER.

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

5 The Experimental Environment

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

1. a collection of textual document;
2. a set of spoken queries with associated relevance assessments recognized at different levels of WER;
3. an IR system.

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

5.1 The Test Collection

Since to the best of my knowledge there is no test collection available with spoken queries, we had to generate it from an existing textual collection[2].The collection used is the *TREC-5 B* a subset of the collection generated for TREC 5 [13]. The collection is made of selected full text articles of the Wall Street Journal (years 1990-92). Some of the characteristics of this test collection are reported in table 1.

An example of a document of the WSJ collection is reported in the following. Notice that documents are marked up using SGML.

```
<DOC>
<DOCNO>
WSJ900402-0192
</DOCNO>
<DOCID>
900402-0192.
</DOCID>
<HL>
   VLSI to Post Profit
   Matching Forecasts
   For the First Quarter
</HL>
```

[2] At ICSI we are currently generating a collection that will have spoken queries and spoken documents. This collection will enable more complex experimentations and analysis for SDR and SQR.

Data sets:	*WSJ 1990-92*
num. of documents	74.520
size in MB	247
num. of queries	50
unique terms in documents	123.852
unique terms in queries	3.504
avg. document length	550
avg. document 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 relevant doc. per query	30

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

```
<DATE>
04/02/90
</DATE>
<SO>
WALL STREET JOURNAL (J), PAGE A8B
</SO>
<CO>
VLSI
</CO>
<IN>
DOW JONES INTERVIEW (CEO)
</IN>
<LP>
   NEW YORK -- VLSI Technology Inc.'s first-quarter earnings
should meet analysts' expectations, the company's chairman
and chief executive officer, Alfred J. Stein, said.
   "We expect to do as well as the analysts are projecting .
. . between five and eight cents a share," Mr. Stein added.
VLSI makes standard and customized integrated circuits.
</LP>
<TEXT>
   Mr. Stein noted that late last year, the company guided
analysts' first-quarter projections lower from earlier
estimates of around 15 cents a share.
   He said a slowdown in standard chip-set sales and a drop
in demand for custom chips by its largest customer, Apple
Computer Corp., stalled revenue growth in the quarter.
   The company had a loss of $6.3 million in the first
quarter of 1989, largely due to problems during the start-up
of its chip plant in San Antonio, Texas. VLSI earned 11 cents
a share in the latest fourth quarter.
   The company expects to release its first-quarter earnings
```

```
April 12.
   Mr. Stein said seasonal slowdown in Far Eastern demand for
the chip sets was partly responsible for damping growth in
the first quarter.
   The region's IBM-compatible computer makers use the sets
in personal computers that see their strongest sales before
the Christmas holidays. Far Eastern demand generally slacks
off in the first quarter, Mr. Stein said. Shipments to the
Far East account for about half of the company's chip-set
sales.
   Mr. Stein said demand for customized chips by Apple
Computer has recovered from a drop that also depressed
first-quarter revenue.
   "Apple is coming back very strongly to us," Mr. Stein
asserted. He added, however, that the impact of Apple's
resumed demand won't be felt until the second and third
quarters of this year.
   Apple accounted for 13% of VLSI's revenue in 1989, while
sales to International Business Machines Corp. rose to about
10% of total revenue.
   Mr. Stein said sales to IBM will exceed sales to Apple
this year because of increasing shipments to IBM, not because
of shrinking sales to Apple. The increasing importance of IBM
as a customer illustrates VLSI's strategic shift toward sales
of "application-specific standard product," primarily
standardized chip sets for personal computers, over the
customized chips designed for Apple Computer and others.
</TEXT>
</DOC>
```

In this work I 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.

I also used a set of 35 queries (or topics, as they are called in TREC) with the corresponding set of relevant documents. These queries were originally in textual form. An example of one of the queries follows:

```
<top>

<head> Tipster Topic Description

<num> Number: 102

<dom> Domain: Science and Technology

<title> Topic: Laser Research Applicable to the U.S.'s Strategic Defense
Initiative

<desc> Description:
```

```
Document will report on laser research related, or potentially related, to
the U.S.'s Strategic Defense Initiative.

<smry> Summary:

Document will report on laser research related, or potentially related, to
the U.S.'s Strategic Defense Initiative.

<narr> Narrative:

A relevant document will provide information on laser research which is
either explicitly linked to the U.S.'s Strategic Defense Initiative (SDI,
also known as "star wars") or which has potential applicability to the
development of SDI laser technology.  Potentially relevant research could be
referred to as research into directed energy weapons, or high energy lasers,
or, more generally, referred to as military laser research.  A relevant
document could take many forms -- award of contract, report of research
results, or abstract of a professional paper on lasers.  However, a document
clearly focused on use of low-power lasers in consumer products, surgical
instruments, or industrial cutting tools is NOT relevant.

<con> Concept(s):

1. Strategic Defense Initiative, SDI, star wars, peace shield

2. laser, directed energy weapon, high energy weapon

<fac> Factor(s):

<nat> Nationality: U.S.

</fac>

<def> Definition(s):

</top>
```

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 query short enough to be a more realistic example of a “real” user query.

5.2 Spoken Queries

The original queries described in the previous section were originally in textual form. It was necessary to produce then in spoken form and have then recognised by a SR system. This work was carried for me by Jim Barnett and Stephen Anderson from Dragon Systems Inc.

	b10	*b40*	*b320*	*t120*	*t160*	*t200*	*t240*	*t320*
Avg Subs %	35.5	24.2	18.8	49.8	31.5	22.7	20.0	19.1
Avg Del %	4.2	2.6	2.6	2.9	3.0	2.6	2.6	2.6
Avg Ins %	11.3	7.8	6.0	21.8	12.4	8.4	6.6	6.0
Avg Err %	*51.0*	*34.6*	*27.4*	*74.5*	*46.8*	*33.6*	*29.2*	*27.7*
Avg Sent Err %	56.7	47.0	39.1	66.5	51.3	42.2	40.4	40.0

Table 2: Characteristics of the different query sets.

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 taht has a 20.000 vocabulary and a bigram language model trained on the Wall Street Journal data. By altering the width of the beam search[3], 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 will not be presented. Different sets of transcripts for the same query set were generated. The characteristics of these sets of transcripts (I will refer to them as query sets) are reported in table 2.

Since the names used by Barnett and Anderson to identify the sets are meaningless if the process used to generated them is not explained and since this is not important for my analysis, I will denote each set with a name that refers to the rounded average WER of that set (i.e. the *AvgErr* in table 2), for example the set *b40* will be refered to henceforth as 35. The set identified by 0 (not reported in the table) is the perfect transcript. An example of the what these error rates actually means on the text of the queries is reported in the appendix.

5.3 The Information Retrieval System

The system used in the context of the work reported in this paper is an experimental IR toolkit developed at Glasgow University mostly by Mark Sanderson [21]. 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 the such a modular architecture. This system was used in the experiments for the Glasgow IR group submissions to TREC-4 [5], TREC-5 [22], and TREC-6 [6].

This system was chosen as the IR platform for the experiments reported in this paper because it implemented a model based on the classical $tf - idf$ weighting

[3] The beam width was chosen as the major parameter to alter because Barnett and Anderson believed that this yields relatively realistic recognition errors.

schema.

The $tf - idf$ weighting schema works as follows. Terms found in documents and queries have first their case normalised, then, any of these terms appearing in a stoplist are removed. The remaining terms are suffix stripped using the Porter stemming algorithm [19]. The remaining document terms are weighted using the $tf - idf$ weighting scheme. 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 in [9]:

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

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 $tf - idf$ retrieval strategy simply evaluates the RSV of a document with respect to a query as the product of the two weighting components and ranks the documents in the collection based on a score. The score for each document is calculated by summing the $tf - idf$ weights of any query terms found in that document:

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

In the IR literature there exists many variations of this formula depending on the way the tf and idf weights are computed [20]. I chose this one because it is the weighting scheme I am most familiar with. Other weighting schemes may prove to be more effective.

5.4 The Evaluation Methodology

The evaluation was performed according to the current practice for IR systems and in accordance with the evaluation methodology used in the various TREC conferences.

The main IR effectiveness measures are Recall and Precision. *Recall* (R) is defined as the proportion of all documents in the collections that are relevant to a query and that are actually retrieved. *Precision* (P) is the proportion of the retrieved set of documents that is also 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, he 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 with the different query sets and precision and recall values were evaluated. The results reported in the following tables and graphs are averaged over the entire sets of 35 queries.

6 Effects of Word Recognition Errors in Spoken Queries on the Effectiveness of an IR System

This section reports some of the results of the experimental analysis of the effects of WRE in SQR. Not all the results of the experiments carried out are presented; only the most interesting ones.

6.1 Effects of WRE on the Standard IR System Configuration

The first analysis was directed towards studying the effects of different WERs in spoken queries on the effectiveness of a 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 that can be removed from the IR indexes without loss of effectiveness. Terms 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 [30, pag.18] and the stemming algorithm commonly known as "Porter algorithm" were used [19]. Figure 2 depicts the results of WRE at different WER in queries over the effectiveness of the IR system using the above standard text-based configuration.

It can be noted that the best results are obtained 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 reasons for this behavior are due to the fact that the IR system parameters are not optimized for the collection and no precision enhancement technique, like for example the use of noun phrases, are employed in the experimentation reported in this paper. 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 and would not allow an easy analysis of the causes of the loss of effectiveness.

Figure 2 also shows that for WERs ranging from 27% to 35% there is not much difference in effectiveness. The little difference that can be seen in the figure is not

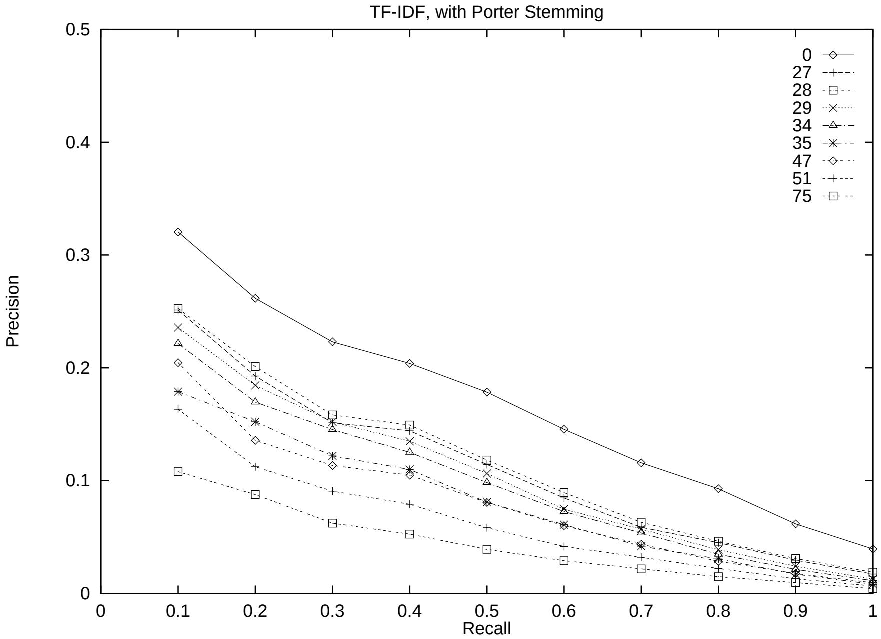

Figure 2: Results using the $tf - idf$ weighting scheme and the Porter stemming.

statistically significant. Moreover, some higher levels of WER seems to do better that lower ones; this again is not statistically significant. Serious loss of effectiveness can only be observed at over 50% WER. We can then conclude that the standard IR is quite robust to WER in spoken queries.

6.2 Effects of WRE on Different IR System Configurations

In order to study the effects of WRE on the effectiveness of SQR, a 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 SQR 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 SQR, 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 the carried out to study this effect, looking at single query terms. 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 weighting scheme. The frequency of occurrence

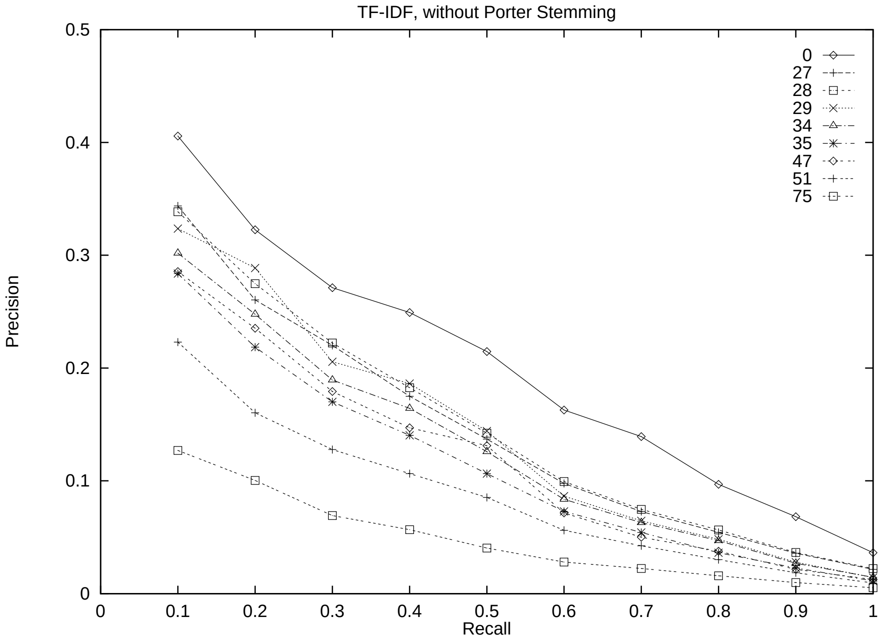

Figure 3: Results using the $tf - idf$ weighting scheme and no stemming.

of a term a document is not considered. It is again surprising to observed that the *idf* weighting scheme produces the same level of effectiveness than $tf - idf$. This is in contrast to what generally happens in textual IR, where term within document frequency is important information for the weighting scheme [11]. Moreover, figure 5 confirms that the use of stemming is detrimental to the effectiveness of an IR system in SQR, as already observed previously in figure 3.

Other experiments involving the use of different versions of the *tf* weighting scheme (the tf_{10}, for example) and of different sizes of stoplists did not produce significantly different results from the one reported here. For this reason, the data obtained from those experiments will not be presented.

From the above data we can conclude that results taking for granted in textual IR, like the effectiveness of weighting schemes based on some variation of the $tf - idf$ weighting scheme, or the use of stemming, do not seem to work well in SQR. The motivation for this behavior are not easy to find and more experimentation is surely needed before making any dangerous generalization.

Table 3 reports the best results obtained in the experimentation, which refers to 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, except at the lowest levels of recall. 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 con-

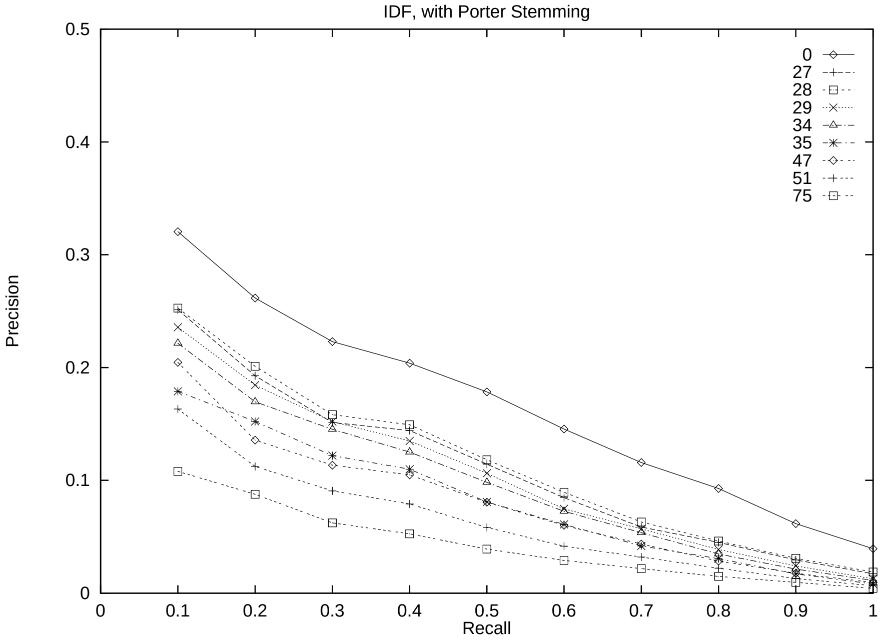

Figure 4: Results using the *idf* weighting scheme and the Porter stemming algorithm.

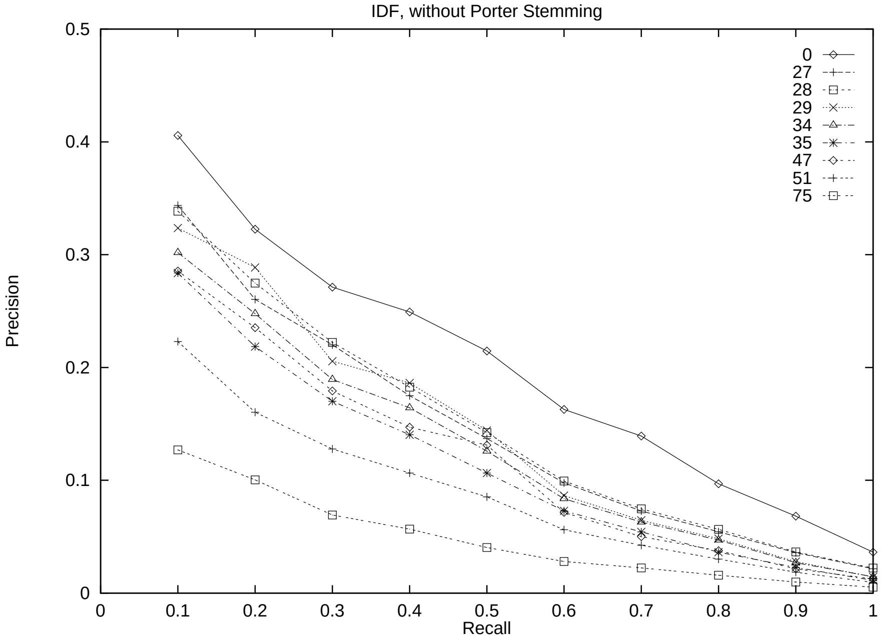

Figure 5: Results using the *idf* weighting scheme and no stemming.

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.

siderably 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 function, that is non content-bearing words. 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 degradation in the effectiveness of the IR process.

6.3 Effects of WRE in Relation to 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 detrimental effects on short queries than on long ones. A 50% WER means that on average half of the terms in a query are missrecognized. 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 10 missrecognized 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 and figure 6 report 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

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.

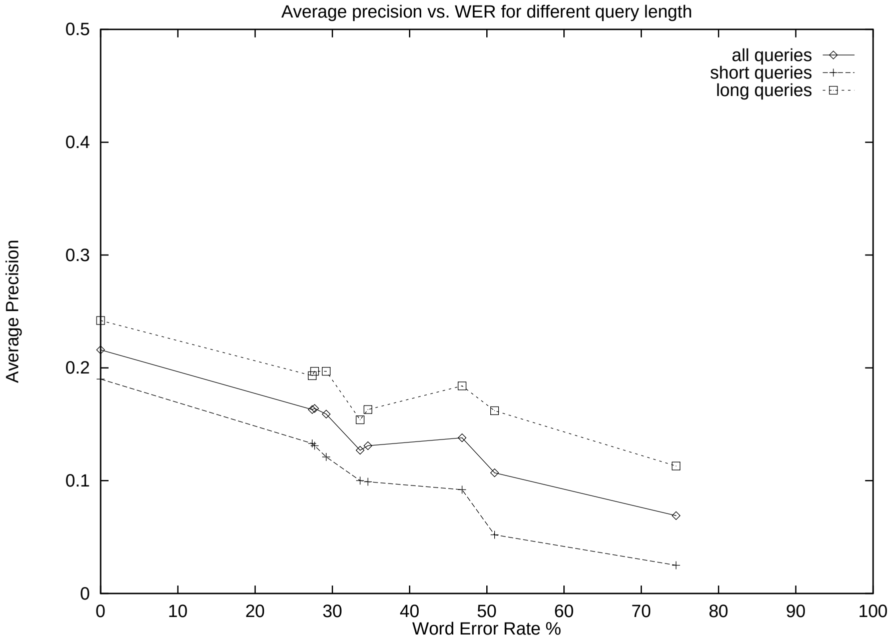

Figure 6: Relation between word recognition errors and average precision by query length.

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.

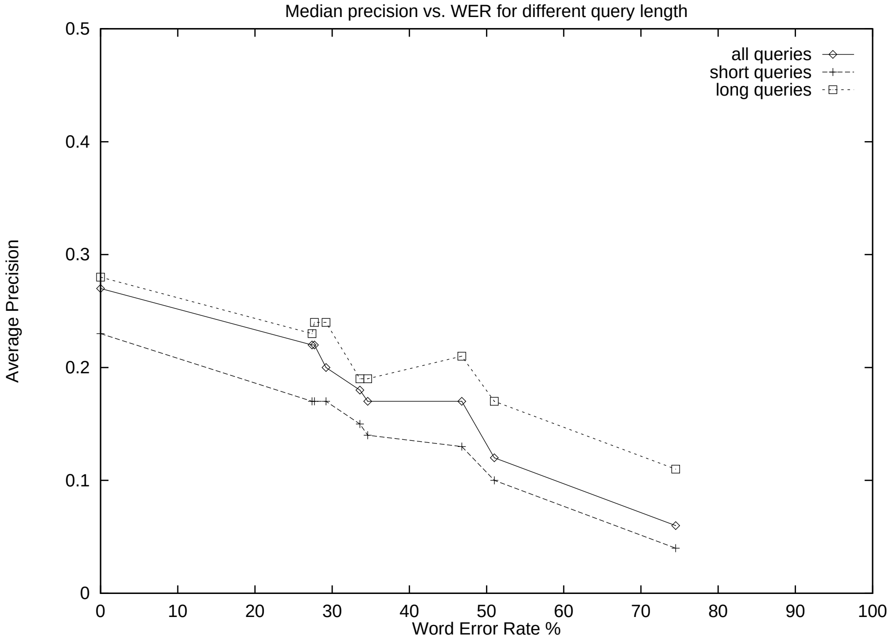

Figure 7: Relation between word recognition errors and median precision by query length.

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 give better performance than some lower WERs, can be explained by the correct recognition of one or more important terms than enabled to find one or more relevant documents than queries at lower levels of WER. This event should be consider not uncommon and can only be ruled out by experiments with larger sets of queries.

Table 5 and figure 7 report 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 this queries. Next section provide some indications of what techniques could be used to improve the performance of an IR system in SQR.

7 Possible Techniques of Recovery from Word Recognition Errors in Spoken Queries

Given the acceptable level of effectiveness of an IR system performing SQR at levels of WER roughly below 40%, we can conclude that it will be quite likely that in the first 10 retrieved documents 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 short spoken summary [28]. 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 me to conclude that *relevance feedback* [12] could be a good strategy to improve effectiveness in a SQR 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. This approach will be tested experimentally in the future.

Another important 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 [18]), and there exist already a number of techniques that we might be able to use in this context [27]. This too will be the object of future work.

8 Conclusions and Future Work

This paper reports on an experimental study on the effects of WRE on the effectiveness of a IR system performing SQR. The results show that classical IR techniques are quite robust to considerably high levels of WER (in the range 27-47%), 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, because of the noise in telephone communication. 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 relevance feedback as a way of recovering 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. I 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.

A Example of the Transcripts Used in the Experimentation

The following examples give an idea of the different quality of the SR system transcripts corresponding to different average WER. The query reported is topic 102, the same reported in section 5.1.

0: laser research applicable to the u. s.'s strategic defense initiative document will report on laser research related or potentially related to the u. s.'s strategic defense initiative strategic defense initiative s. d. i. star wars peace shield laser directed energy weapon high energy weapon

27: a laser research applicable to the u. s.'s strategic defense initiative doctor that will report to own laser research related or potentially routed thro ugh u. s.'s strategic defense initiative strategic defense initiative s. t. i. star wars peace shield moisture directed energy revco high energy will come

28: a laser research applicable to the u. s.'s strategic defense initiative doctor that will report to own laser research related or potentially routed thro ugh u. s.'s strategic defense initiative strategic defense initiative s. t. i. star wars peace shield moisture directed energy revco high energy will come

29: a laser research applicable to the u. s.'s strategic defense initiative doctor that will report to own laser research related or potentially routed thro ugh u. s.'s strategic defense initiative strategic defense initiative s. t. i. star wars peace shield moisture directed energy revco high energy will come

34: a laser research applicable to the u. s.'s strategic defense initiative doctor that will report to own laser research related or potentially routed thro ugh u. s.'s strategic defense initiative strategic defense initiative s. t. i. star wars peace shield laser directed energy revco high energy will come

35: placer research applicable to the u. s.'s strategic defense initiative doctor that will report the only is a research related or potentially routed thr ough u. s.'s strategic defense initiative strategic defense initiative s. t. buy star wars peace shield lawyers are directed energy program high and truck them

47: placer we stirred up a couple to us his strategic offensive initiative doctor them over pork always a research related or potentially routed through us his strategic offensive a shipment strategic defense initiative s. t. i. star wars peace shield winter directed energy revco high energy will come

51: the ways research up a couple to the u. s.'s strategic defense initiative thought command over pork always research related potentially removed the u. s.' s strategic defense initiative strategic defense initiative s. p. five star war us peace shield boys their directed energy program high and truck on

75: placer we certificate the poll to us this strategic offensive initiative botha met with pork a laser research related upper tension beirut to us whose st rategic offensive a shipment strategic the fed submission to s. t. on star wars peace shield windsor there could energy weapon high and to a pub

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, November 1997.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo. Experiments in spoken queries for document retrieval. In *Eurospeech 97*, volume 3, pages 1323–1326, Rodhes, Greece, September 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, August 1999.

[5] F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. Experimenting retrieval by logical imaging on TREC. In *Proceedings of the TREC Conference*, pages 509–525, Washington D.C., USA, November 1995.

[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, November 1998.

[7] S. Deerwester, S.T. Dumais, G.W. 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.B. Frakes. Stemming algorithms. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 8. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

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

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

[11] D. Harman. Ranking algorithms. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 14. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[12] D. Harman. Relevance feedback and other query modification techniques. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 11. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

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

[14] J.N. Holmes. *Speech synthesis and recognition.* Chapman and Hall, London, UK, 1988.

[15] M. Magennis and C.J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of ACM SIGIR*, pages 324–332, Philadelphia, PA, USA, July 1997.

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

[17] E. Mittendorf and P. Schauble. Measuring the effects of data corruption on Information Retrieval. In *Proceedings of the SDAIR 96 Conference*, pages 179–189, Las Vegas, NV, USA, April 1996.

[18] 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, February 1991.

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

[20] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, May 1976.

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

[22] M. Sanderson and I. Ruthven. Report of the glasgow ir group submission. In *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, Washington D.C., USA, November 1996.

[23] M.A. Siegler, M.J. Witbrock, S.T. Slattery, K. Seymore, R.E. Jones, and A.G. Hauptmann. Experiments in spoken document retrieval at CMU. In *Proceedings of TREC-6*, Gaithersburg, MD, USA, November 1997.

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

[25] A. Singhal, J. Choi, D. Hindle, and F. Pereira. AT&T at TREC-6: SDR Track. In *Proceedings of TREC-6*, Washington DC, USA, November 1997.

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

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

[28] 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.

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

[30] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[31] E. Voorhees, J. Garofolo, and K. Spark Jones. The TREC-6 spoken document retrieval track. In *TREC-6 notebook*, pages 167–170. NIST, Gaithersburgh, MD,USA, 1997.

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

[33] 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.